Grounding Criterion: Toward a Formal Theory of Grounding

Tim Paek
Eric Horvitz

April 6, 2000

Technical Report
MSR-TR-2000-40

Microsoft Research
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052

Abstract

In a conversation, participants establish and maintain their mutual belief that their utterances have been understood well enough for current purposes – a process that has been referred to as *grounding*. In order to make a *contribution* to the conversation, participants typically do more than just produce the right utterance at the right time; they coordinate the presentation and acceptance of their utterances until they have reached a sufficient level of mutual understanding to move on, a level defined by the *grounding criterion*. Recent interest in employing grounding for use in collaborative dialog systems has highlighted difficulties in rendering hitherto qualitative intuitions about grounding into formal terms. In this paper, we propose a formalization of grounding based on decision theory that captures key intuitions about the contribution model while providing an explicit method for determining the grounding criterion. We illustrate the formalization by reviewing interactions between a user and a prototype spoken dialog system called the *Bayesian Receptionist*.

Introduction

When people engage in a conversation, they do so with the intent of making themselves understood to all the participants. To do this, they need to assure themselves, as they produce each utterance, that the others are at the same time attending to, hearing, and understanding what they are saying. This involves coordinating not only the content of their utterances, but also the process. Hence, participants not only produce an utterance, they continually monitor other participants to make sure their utterances were heard and understood well enough for current purposes (Clark & Schaefer, 1987, 1989).

An example of this process is feedback. When participants in a conversation understand each other, they signal this through acknowledgements, such as "uh huh," or by simply moving on in the conversation. On the other hand, when they are uncertain about their own understanding or that of another participant, they signal this by attempting to repair the situation. Speakers look for signs of understanding from their listeners. In fact, they often repeat themselves if listeners do not provide adequate feedback of attention (Clark, 1996).

Conversation from this perspective may be best regarded as a collaborative effort, a type of joint activity, in which participants coordinate the presentation and acceptance of their utterances to establish, maintain, and confirm mutual understanding (Clark, 1996; Cohen & Levesque, 1994; Grosz & Sidner, 1990). The process by which they do this has been called *grounding* (Clark & Brennan, 1991; Clark & Schaefer, 1987, 1989; Clark & Wilkes-Gibbs, 1990). Refraining from grounding results in communication failure, the repair of which may be costly both in terms of time and effort (Brennan & Hulteen, 1995; Hirst et al., 1994; Horvitz & Paek, 1999; Paek & Horvitz, 1999; Traum & Dillenbourg, 1996).

In order to make a *contribution* to a conversation, participants must not only convey or specify some content, they must collaboratively try to establish the mutual belief that the listeners have understood what the speakers meant as part of their shared set of beliefs or *common ground*. In this view, conversations do not proceed utterance by utterance, but contribution by contribution.

While researchers have long deliberated various ways of representing common ground or similar notions of mutual belief (Cohen & Levesque, 1991, 1994; Grosz & Kraus, 1993; Haddadi, 1995; Halpern & Moses, 1990; Heeman & Hirst, 1995), relatively little work has focused on the process of grounding (Brennan & Hulteen, 1995; Traum, 1999). With rising interest in building collaborative dialog systems that make use of recent advances in spoken language technology, a growing number of researchers have begun to explore the suitability of grounding as a design model for human–computer collaboration

(Horvitz, 1999; Traum, 1999). It is easy enough to say that participants in a conversation seek out and establish mutual belief of understanding, but the task of formalizing it so that it can be useful for building and evaluating collaborative systems is very challenging.

The most immediate difficulty is that of identifying when mutual understanding has been *sufficiently* reached before moving on to the next contribution. Determining a sufficient level of belief depends on the *grounding criterion*: "The contributor and the partners mutually believe that the partners have understood what the contributor meant to a criterion sufficient for current purposes" (Clark & Schaefer, 1989). As researchers have shown, the grounding criterion varies widely depending on the goals and features of the joint activity, such as communication medium and time criticality (Clark & Brennan, 1991). For example, if a speaker is relating a secret code to deactivate a bomb, a simple "uh huh" from the listener will hardly suffice. Feedback that increases the belief of mutual understanding, such as reiteration of the code by the listener, is needed. Listeners may also want to assure themselves of their own beliefs in what they heard by explicitly confirming the secret code.

As the example demonstrates, defining a sufficient level of mutual belief requires at least two critical components for any formalization: first, the quantification of belief, and second, a consideration of how "sufficient" belief changes depending on the stakes of pursuing various actions in different contexts. We now present a formalization that integrates these two components using decision theory.

Grounding as Decision Making

The approach we take is to treat the process of grounding as decision making under uncertainty. In a conversation, uncertainty abounds; this includes uncertainty about what words were uttered, what intentions a speaker had in producing those words, and how well the listener understood those words and intentions. Real-time decisions need to be made in the face of this uncertainty. In grounding, such decisions involve communicative actions that may or may not improve mutual beliefs about understanding. For example, listeners may ponder whether to provide immediate feedback of their own lack of understanding or wait to see if their confusion is eventually cleared up. Actions have consequences; hence, we assign utilities to outcomes that ensue from making particular decisions. In decision theory, uncertainty, actions, and utilities all come together within a unifying mathematical framework.

The approach we take is similar to recent work in applying reinforcement learning to dialog (Levin et. al, 1997; Levin et. al, 2000; Singh et. al, 1999; Walker et. al, 1998). There, dialog is construed as a Markov Decision Process (MDP). Treating utterances as observations and using a reward or utility function for arriving at various outcomes, the goal is to derive a sequence of action that maximizes the reward. While the formalization presented here follows in the same tradition as the MDP framework, that should not detract from the focus of the approach which is on *grounding*, and in particular, on *mutual beliefs* of understanding. The goal here is to use the language and machinery of decision theory to formalize hitherto qualitative intuitions about mutual beliefs and how they are influenced by the stakes of making a particular decision.

We now present a first set of definitions. Later we build on these definitions to handle more complex cases. Finally, we illustrate the formalization by reviewing interactions between a user and a prototype spoken dialog system called the *Bayesian Receptionist*.

1 Formalizing Grounding

In the collaborative view of conversation, speakers do not produce an utterance in isolation. They do so with the intent of making themselves understood. Hence, they design an utterance for their audience, and

thereafter, seek out signs of understanding. Similarly, listeners do no passively receive utterances. They attempt to understand what the speakers meant, and respond accordingly. Every response, even silence, which relinquishes the opportunity to repair, constitutes feedback or evidence of understanding. We now examine the two perspectives in formal terms.

1.1 Listeners

After speakers produce an utterance, listeners can take a myriad of actions. For now, consider that listeners resolve these actions into one of two choices: either to engage or not engage in a repair. Denote this as R and ¬R. Furthermore, consider that listeners can either comprehend or not comprehend what was meant by the utterance. Denote this as C and ¬C. Since listeners comprehend to varying degrees between C and ¬C, let the probability $p(C|E)$ denote the likelihood that they comprehended what was meant given all evidence E so far. Intuitively, this likelihood expresses belief on the part of listeners in their own comprehension if they were to answer how well they understood an utterance.

In decision theory, the choice of whether or not to take an action is guided by expected utility (Howard, 1970). We can identify thresholds for action by considering four deterministic outcomes and the uncertainties in each (Horvitz, 1990). The listener either comprehends or does not comprehend the utterance, and for each of these states, the listener either engages or does not engage in a repair. Mapping an associated value, or *utility*, to each of the outcomes, we obtain:

- $u(C,R)$: the utility of repairing when the listener comprehends the utterance
- $u(C,\neg R)$: the utility of not repairing when the listener comprehends
- $u(\neg C,R)$: the utility of repairing when the listener does not comprehend the utterance
- $u(\neg C,\neg R)$: the utility of not repairing when the listener does not comprehend

A partial ordering for the four utilities emerges upon reflection. Intuitively, it seems that not repairing when the listener comprehends is better than repairing, and likewise, that repairing when the listener does not comprehend is better than not repairing. In short,

$$u(C,\neg R) \geq u(C,R) \; ; \; u(\neg C,R) \geq u(\neg C,\neg R)$$

With this partial ordering, the expected utility of engaging in a repair sequence is computed as the sum of the utilities for each outcome with a repair, weighted by its likelihood.

$$eu(R)=p(C|E)u(C,R)+[1-p(C|E)]u(\neg C,R)$$

The expected utility of not engaging in a repair follows analogously.

$$eu(\neg R)=p(C|E)u(C,\neg R)+[1-p(C|E)]u(\neg C,\neg R)$$

By plotting the expected utility as a function of the likelihood of understanding, as shown in Figure 1, we arrive at the following definitions.

Definition 1 *From the listener's perspective, the point at which the expected value of engaging in a repair is equal to the expected value of not engaging is* p^*.

Definition 2 *When condition* $p(C|E) > p^*$ *holds, it can be said that the listener comprehends the utterance sufficiently to move on.*

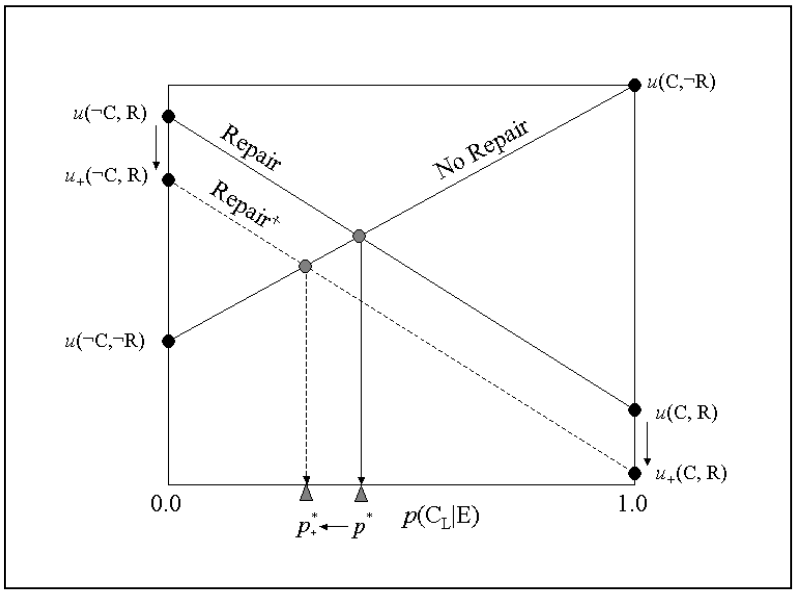

Figure 1. The expected utility of repair versus no repair yielding a probability threshold. Time dependent utilities for repair lower the threshold.

As shown in Figure 1, if the listener's probability of comprehension, $p(C|E)$, falls short of the threshold p^*, the action with the highest expected utility is to repair. On the other, if it exceeds p^*, the optimal action is not to repair.

Given that sufficient belief hinges on p^*, Figure 1 also illustrates how belief can be influenced by the stakes involved in taking action and arriving at various outcomes. Suppose that the utility of repair decreases with time whether or not the listener comprehends, or alternatively, that the cost of repair increases with delay. This is depicted by dropping the utility line for repair. From the graph, it is apparent that with delay, the cost of repair increases, causing p^* to decrease to p^*_+. Intuitively, this makes sense since in cases where the cost of delayed action is increasing, as for instance under time pressure, listeners require less certainty in comprehending an utterance.

It is important to note that utilities can fluctuate, just as in the case above. Furthermore, while the four utilities mentioned previously were chosen since they are relatively easy to assess, a system may fit its own utility function based on data rather than assuming simple linearity. The expected utility equations remain the same.

1.2 Speakers

After speakers produce an utterance, they monitor and seek out evidence of understanding from their listeners. If listeners respond by displaying gestures indicative of confusion, such as a scowl or even worse, a repair, that is strong evidence that p^* has not been reached. On the other hand, gestures indicative of sufficient understanding, such as not repairing, may be more subtle. Indeed, researchers have assigned different evidential strength to these various gestures (Clark & Schaeffer, 1989).

For speakers, the primary decision to make after receiving response evidence *e* is whether or not to move on; that is, to accept that the listener comprehended what they meant sufficiently to move on to the next

utterance. Denote this as M and ¬M. Since speakers only have an estimate of the true comprehension of listeners, let $p_S(C \mid E, e)$ denote the speaker's belief in the listener's comprehension of the utterance.

Following the same scheme as before, we consider that moving on when the listener comprehends the utterance is better than not moving on, and likewise, that not moving on when the listener does not comprehend is better than moving on. This intuition results in the partial ordering:

$$u(C,M) \geq u(C,\neg M) \; ; \; u(\neg C,\neg M) \geq u(\neg C,M)$$

The following definitions now apply:

Definition 3 *From the speaker's perspective, the point at which the expected value of moving on is equal to the expected value of not moving on is* p_S^*.

Definition 4 *When the condition* $p_S(C \mid E, e) > p_S^*$ *holds, it can be said that the speaker believes the listener to have understood some utterance sufficiently to move on.*

Oftentimes, speakers take their p_S^* to be same as the listener's p^*, or p_L^*, to keep notational consistency. This is due to the fact that they design their utterances for the listeners. When speakers do this, they keep in mind what they share with the listeners in common ground, including awareness of how beliefs can change given the stakes involved in understanding an utterance.

An example of this process is the use of "instalments" (Clark & Brennan, 1991; Clark & Schaefer, 1989; Clark & Wilkes-Gibbs, 1990). When speakers have complicated information to present, they often transmit it in small portions to make sure each component is fully understood. For example, in the earlier bomb scenario, speakers there will provide the deactivation code number by number, and seek explicit confirmation of each instalment.

The formalization we present sheds light on how instalments work. From assessing the tremendous stake in moving on when the speaker does not comprehend the code, the speaker derives a high p_S^* for sufficiently understanding the utterance. The speaker knows that listeners are more likely to understand a single number than the entire code at once; in other words, $p_S(C \mid E, e)$ is likely to be higher for single numbers. Hence, to improve the odds that $p_S(C \mid E, e) > p_S^*$, they give the code in instalments. Listeners, likewise, know the stakes and have a high p_L^*. Hence, to boost their belief in their own comprehension $p_L(C \mid E)$, they engage in confirmations of each number.

1.3 Sufficient Mutual Understanding

As explained previously, to make a *contribution* to a conversation, participants must not only specify some content, they must ground it. We are now in a position to define a contribution.

Definition 5 *A contribution has been made with respect to some utterance when the listener sufficiently comprehends the utterance and when the speaker believes the listener to have comprehended the utterance. In short, a contribution has been made when the following statement holds true:*

$$[\, p_S(C \mid E, e) > p_S^* \,] \wedge [\, p_L(C \mid E) > p_L^* \,]$$

Consistent with the original Clark & Schaefer contribution model, we can say that participants reach *sufficient mutual understanding* of their utterances when they have made a contribution. Until then, the process of making a contribution with respect to some content is called *grounding*.

Given that participants identify p_S^* and p_L^* by considering the stakes, grounding is equivalent to belief updating. In other words, participants continually monitor each other to update their $p_S(C \mid E, e)$ and $p_L(C \mid E)$ until a contribution is made.

Definition 5 also allows for the possibility that p_S^* may not be the same as p_L^*. Generally, speakers assume that listeners share the same belief about how much understanding is sufficient. However, the two thresholds may be quite distinct. For example, students may feel they have grasped what the teacher meant sufficiently to move on, but the teacher may not. In such a case, a contribution has not been made and the participant must continue to ground the intended lecture.

1.4 Higher Order Beliefs

So far, we have only dealt with first order beliefs. However, the formalization allows there to be higher order beliefs: that is, beliefs about beliefs.

Definition 5 can be viewed in several perspectives. For outside observers trying to identify when participants have made a contribution, four beliefs are needed: $p_S(C \mid E, e)$, p_S^*, $p_L(C \mid E)$, and p_L^*. For listeners, they already introspectively know $p_L(C \mid E)$ and p_L^*.; hence, they only need two beliefs: $p_S(C \mid E, e)$ and p_S^*. Likewise, speakers only need to know $p_L(C \mid E)$ and p_L^*.

Whatever perspective is taken, the needed beliefs are themselves uncertainties. A system taking that perspective may want to have beliefs about these uncertainties. This can be done by maintaining a probability distribution over that belief and calculating its expectation. For example, a system acting as a speaker can estimate the listener's comprehension $p_L(C \mid E)$ using:

$$e_S[p_L(C \mid E)] = \int_{p_L} p_S(p_L(C \mid E))dp_L$$

An alternative to the expectation is to embed beliefs. For example, a system acting as a speaker can decompose C as follows:

$$\mathrm{C} \sim [\, p_L(C \mid E) \succ p_L^* \,]$$

Rewriting $p_S(C \mid E, e)$, we obtain:

$$[\, p_S(p_L(C \mid E) \succ p_L^* \mid E, e) \succ p_S^* \,]$$

This embedded belief explicitly represents at the assumption that speakers take their p_S^* to be the same as p_L^*, even though these two thresholds may differ drastically, as previously discussed.

We discuss elsewhere (Anonymous, 2000) the merits of using embedded beliefs and expected beliefs, as well as how to handle infinite orders of beliefs.

2 Dialog System Example

To further concretize the formalization, we describe how researchers have implemented it in a spoken dialog system. The *Bayesian Receptionist* (Horvitz & Paek, 1999; Paek & Horvitz, 1999) employs probabilistic networks, natural language parsing, and speech recognition to guide conversations about tasks typically handled by front desk receptionists at the Microsoft campus. The system assumes the perspective of listeners in not only maintaining models of its own comprehension but that of the speaker, or user, as well.

When a user first interacts with the system, all available evidence is used to assess its belief about what the user wants. This constitutes its $p_L(C \mid E)$. It gathers evidence by recognizing an initial utterance, parsing it, and observing words, as well as semantic and syntactic components in a belief network. A *belief network* provides an efficient way of encoding probabilistic dependencies between random variables using a directed acyclic graph (Pearl, 1991). Extensive observational and questionnaire studies went into identifying the relevant variables for the network as well as the probabilistic dependencies (Horvitz & Paek, 1999).

Once the system infers its beliefs about what the user wants, which is displayed in Figure 2 as a probability distribution, it has to decide what to do. It takes action only when it believes it has comprehended what the user meant sufficiently to do so. The primary action is to provide a service corresponding to what the user wants, such as calling a shuttle. It decides to provide the service only if the goal with the highest probability exceeds Pstar (as in p_L^*). Pstar can be evaluated by considering the utility of providing or not providing a service when the system has correctly or incorrectly understood the goal. In Figure 2, the goal of SHUTTLE has the highest probability (.458), but that is not enough to surpass Pstar.

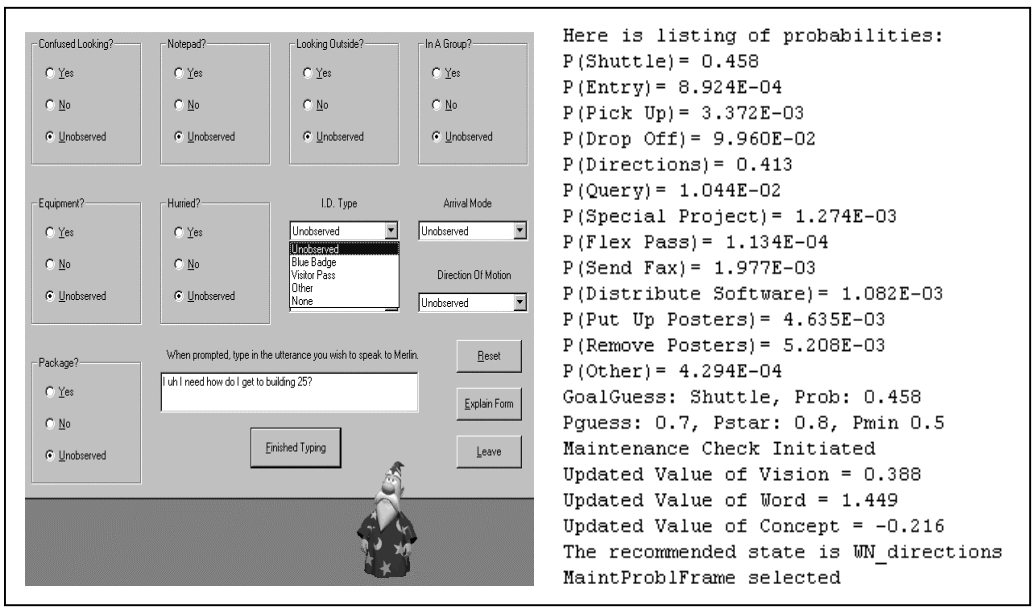

Figure 2. The *Bayesian Receptionist* inferring the goals of the user from an initial utterance and determining if that exceeds its probability threshold for providing a service.

If that fails, the system checks to see if the goal exceeds Pguess, the threshold probability for confirming its belief (e.g., "Did you want a shuttle?"). Again, Pguess can be evaluated by considering the utility of confirming or not confirming belief when the system has correctly or incorrectly understood the goal. This procedure of checking various thresholds continues until it takes action.

The system continues to take action until $p_L(C \mid E)$ exceeds Pstar, but to know when a contribution has been made, it keeps track of whether the condition $p_S(C \mid E, e) > p_S^*$ holds by maintaining first order belief models of the user. This is done by treating the condition as a node in a belief network and attempting to diagnose the state of that node from evidence.

An example of how this works can be seen if the user should all of a sudden leave the interface during an interaction (Paek & Horvitz, 1999). Here, the system waits for the user to respond to its action. If there is no response, it updates the probability that $p_S(C \mid E, e) > p_S^*$. Eventually, as this probability falls and the belief that the speaker has PREMATURELY TERMINATED the activity (a state in that node) increases, the system assumes that the activity is over without having had the chance to make a contribution with the user.

4 Conclusion

We introduced a formalization based on decision theory that captures key intuitions about grounding while providing an explicit method for determining when sufficient mutual understanding has been reached by participants in a conversation. We then reviewed interactions between a user and a system that utilizes the formalization. We hope that the approach we take will provoke valuable discussion on the nature of grounding and contributions to better assist those exploring the suitability of grounding as a design model for human–computer collaboration.

References

Anonymous. 2000. *Foundations of grounding: Part I. On grounding criterion*. In preparation

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

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

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

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

Clark, H.H. and Schaefer, E.F. (1989) *Contributing to discourse*. Cognitive Science, 13, pp.259-294.

Clark, H.H. and Marshall, C.R. (1981) Definite reference and mutual knowledge. In Elements of Discourse Understanding. Cambridge University Press, pp. 10-63.

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

Cohen, P.R. and Levesque, H.J. (1991) *Teamwork*. Nous, 25/4, pp.487-512.

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

Grosz, B.J. and Kraus, S. (1993). *Collaborative plans for group activities*. In Proceedings IJCAI-93, pp. 367-373.

Grosz, B.J. and 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.

Halpern, J.Y. and Moses, Y. (1990). *Knowledge and common knowledge in a distributed environment*. Journal of the ACM, 37/3, pp. 549-587.

Heeman, P., and Hirst, G. (1995) *Collaborating on referring expressions*. Computational Linguistics, 21/3, pp. 351-382.

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

Horvitz, E. (1990) *Computation and action under bounded resources*. Ph.D. thesis, Stanford University.

Horvitz, E. (1999) *Principles of mixed-initiative user interfaces*. Proc. of CHI '99, ACM Press, pp.159-166.

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

Howard, R. (1970) *Decision analysis: Perspectives on inference, decision, and experimentation*. Proceedings of the IEEE, 58, pp. 632-643.

Levin. E., Pieraccini, R., & Eckert, W. (2000). A stochastic model of human-machine interaction for learning dialog strategies. IEEE Transactions on Speech and Audio Processing, 8, pp. 11-23.

Levin, E., Pieraccini, R., & Eckert, W. (1997). *Learning dialogue strategies within the Markov decision process framework*. In Proc. IEEE Workshop on Automatic Speech Recognition and Understanding

Paek, T. & Horvitz, E. (1999). *Uncertainty, utility, and misunderstanding:* In AAAI Fall Symposium on Psychological Models of Communication, pp.85-92.

Pearl, J. (1991) Probabilistic Reasoning in Intelligent Systems, Morgan Kaufmann Publishers: San Francisco.

Singh, S., Kearns, M., Litman, D., & Walker, M. 1999. *Reinforcement learning for spoken dialogue systems*. In Proceedings of NIPS.

Traum, D. (1999). *Computational models of grounding in collaborative systems*. In AAAI Fall Symposium on Psychological Models of Communication, pp.124-131.

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

Walker, M., Fromer, J., & Narayanan, S. 1998. *Learning optimal dialogue strategies: A case study of a spoken dialogue agent for email*. In Proceedings of COLING/ACL, pp. 1345-1352.