

Contribution Tracking

Contribution Tracking: Models and Skills for Collaborative Language Use Under Uncertainty

David DeVault*
ICT

Matthew Stone**
Rutgers

*This article explores the representations and reasoning by which dialogue systems can work collaboratively and creatively to communicate successfully with interlocutors in situations of transient uncertainty. In these situations, dialogue is shaped by the hierarchy of activities underway in the interaction, by the dynamics of linguistic context, and by the array of possibilities that interlocutors take as open ambiguities. Only by integrating information of all these types can a dialogue agent interpret incremental contributions from interlocutors during periods of uncertainty, or formulate its own context-dependent utterances to help pinpoint the context and resolve ambiguities. An empirical study with an implemented human-computer dialogue system in a referential communication task shows how these skills of **contribution tracking** enable new, more flexible, more robust, and more coherent strategies for interactive language use.*

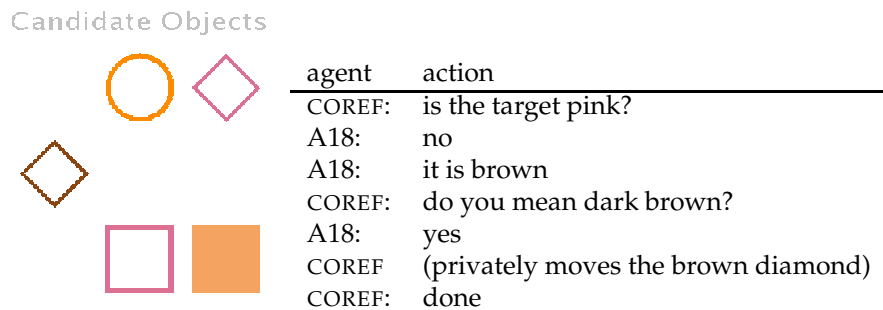
1. Introduction

People can work together to make sure they understand one another. This is both our commonsense understanding of conversation (Clark 1996) and an important finding of empirical investigations (Clark and Wilkes-Gibbs 1986; Clark and Schaefer 1989; Brennan and Clark 1996). Collaboration emerges not only as an essential characteristic of language use but a crucial reason why human communication is so robust to failure. When we do not understand what others say to us, we can leverage our conversational abilities to follow up provisional interpretations of what has been said and eventually arrive at a sufficient understanding. For example, when identifying an object, speakers are prepared to give many alternative descriptions, and listeners not only show whether they understand each description but often help the speaker find one they do understand (Clark and Wilkes-Gibbs 1986).

In computational linguistics too, researchers in pursuit of more natural, more flexible and more robust interaction with conversational agents have sought to improve systems' collaborative conversational skills. The challenge involved can be illustrated with the simple fragment of clarification dialogue shown in Figure 1. The example is drawn from the evaluation experiment we report in Section 4, in which human users interacted by teletype with our implemented agent COREF. As the user attempts to

* Insitute for Creative Technologies, University of Southern California. 13274 Fiji Way Marina del Rey, CA 90292. E-mail: devault@ict.usc.edu

** Department of Computer Science and Center for Cognitive Science, Rutgers, the State University of New Jersey. 110 Frelinghuysen Road, Piscataway NJ 08854-8019. E-mail: matthew.stone@rutgers.edu

**Figure 1**

COREF asks a clarification question to disambiguate reference to a graphical display.

identify a specific object from the graphical display, COREF perceives an ambiguity in how *brown* is to be understood, and decides to clarify with *do you mean dark brown?*

To achieve such interaction in a general way, we argue here, a system needs integrated access to representations that characterize the organization of ongoing activity, the dynamics of linguistic context, and the patterns of uncertainty that play out as interlocutors work to achieve shared understanding. These different information sources account for different aspects of dialogue.

- Conversation is organized purposefully, like all human joint activity (Power 1977; Perrault, Allen, and Cohen 1978; Grosz and Sidner 1986; Litman and Allen 1987; Pollack 1990, 1992; Rich, Sidner, and Lesh 2001). We work step by step, breaking larger activities down into their smaller components, each governed by an overall intention and linked to our broader efforts. In Figure 1, for example, the clarification exchange comes as part of the effort of identifying the color of the object, which in turn contributes to broader tasks of identifying the object itself and taking an action involving that object. Models of collaboration emphasize the diversity of moves through which participants can coherently contribute to extended interactions, and the need to describe them in the general terms of purposeful activity. Heeman and Hirst (1995), for example, in their plan-based model of collaborating on referring expressions, illustrate the many possibilities for proposal, counterproposal, refinement and acknowledgment that may be required of interlocutors, over multiple turns, before they agree. Even the interaction of Figure 1 invites us to couple the dialogue structure induced by moves such as *yes* and *done* with the successful outcome eventually obtained.
- Conversation exploits the distinctive status of utterances as linguistic actions, with effects that update an abstract scoreboard or information state (Lewis 1979; Thomason 1990; Bunt 2000; Matheson, Poesio, and Traum 2000; Ginzburg and Cooper 2004). Utterances go on the record; they also license anaphoric relations, raise or resolve questions, establish or discharge obligations, and so forth. Almost every utterance in the dialogue of Figure 1 draws on contextual information established by the preceding utterances: the elliptical expressions *yes*, *no* and *done*, the pronoun *it*, the presuppositional frame of the question *do you mean*. To generate or

understand such expressions in a general, declarative way, depends on maintaining an evolving representation of linguistic context.

- Problematic situations inevitably involve uncertainty. In the situation of Figure 1, for example, COREF may not always understand the user, but neither is COREF simply ignorant of the user's meaning. Rather, COREF is in a position to maintain a set of plausible hypotheses about the object the user is likely to pick and to reevaluate those hypotheses as additional information about the user's intention comes in. There is good reason to think that doing so can make systems more effective at tracking ambiguity and responding appropriately. For example, maintaining a probability distribution over alternative recognition results can make it easier for a system to combine evidence about user intentions from multiple utterances (Bohus and Rudnicky 2006). It can also help a system to choose whether to clarify user input or proceed with a possibly incorrect interpretation (Roy, Pineau, and Thrun 2000; Horvitz and Paek 2001; Williams and Young 2007).

Simply maintaining these representations is not enough, of course: the system's decision making must be sensitive to all of these different representations simultaneously.

In this article, we present a novel approach to these challenges for representation and reasoning in dialogue agents. Our approach involves a set of problem-solving mechanisms we collectively call **contribution tracking**, which describe the distinctive collaborative processes that are required to interpret incremental contributions from interlocutors and formulate new context-dependent utterances during periods of transient uncertainty. In contribution tracking, the overarching problem remains **collaboration**: to coordinate the execution of an agreed task, drawing on a rich shared background of effective joint strategies for getting things done. Interlocutors' moment-by-moment coordination, as in any collaboration, depends on a fundamental operation of **intention recognition**: determining how an agent's actions, in light of shared background knowledge and expectations, commit them to a determinate contribution to the ongoing activity. We address the distinctive properties of utterances by using grammatical knowledge, including the contextual preconditions and contextual effects of specific linguistic forms, as part of the shared background for intention recognition.

The key innovations in our model follow from our assumption that intention recognition potentially involves uncertainty, so that there may be alternative ways to reconstruct a plausible task contribution from the actions a speaker takes and the available background. To estimate the state of the ongoing activity after a newly-observed action, agents must weigh their prior hypotheses about the collaboration against their possibly ambiguous evidence about the speaker's current intention, and propagate their uncertainty forward. To plan an appropriate collaborative response, agents must assess individual moves for their outcomes in aggregate across the likely possibilities. And finally, to coordinate those responses, agents must synthesize utterances that make desired moves without introducing new ambiguities, no matter what interpretation they might receive in light of the different possible states of the conversation.

Our model is implemented in COREF, a task-oriented dialogue system that collaboratively identifies visual objects with human users. We show empirically that to interact successfully in its domain, COREF does need to work collaboratively to resolve ambiguities, and moreover that our model makes COREF to some degree successful in doing so. At the same time, we highlight qualitative aspects of COREF's behavior that depend

on our new synthesis of probabilistic, linguistic and collaborative reasoning. It enables COREF to understand ambiguous acknowledgments as giving partial information about users' meanings. It enables COREF to synthesize a generative range of coherent followup questions to elicit further information from users in context. And it lets COREF resolve uncertainty flexibly over extended patterns of interaction.

These arguments, both theoretical and empirical, constitute the main contribution of the article. They serve to organize a diverse set of skills for the collaborative negotiation of meaning in terms of the models and representations required to achieve them productively, and thus help to map out the design space for new dialogue representations and architectures.

The organization of the article is as follows. We begin in Section 2 by contrasting the multidimensional representation and reasoning we pursue with the alternative perspectives and assumptions adopted in prior research. We continue in Section 3 by presenting a detailed overview of the approach we advocate. Section 4 describes how we implemented this approach in a specific referential communication task and assessed the system's behavior in interactions with human users; Section 5 refers the results of our experiments to the innovations and limitations of our implementation. We conclude in Section 6 by offering a further roadmap to the problems that remain to be tackled on our approach, and signposting some of the key tradeoffs that may limit the applicability of our results.

2. Situating our Work

The specific problem we address in this article is how to reason about context-dependence in conversation while working collaboratively with an interlocutor to reduce ambiguity and achieve common ground. Every utterance in conversation gets its precise meaning in part through its relationship to what has come before. This applies to the clarificatory utterances interlocutors use to acknowledge, reframe or question others' contributions just as it does to fresh contributions. The distinctive issue with such followups is that they must be formulated for a context about which speaker or addressee may be uncertain. The speaker must be able to assess that addressees will understand and respond helpfully to them no matter what the context might be.

This characterization suggests how our work lies at the intersection of three different approaches to specifying dialogue agency. The first is of course prior models of collaboration, notably (Heeman and Hirst 1995; Rich, Sidner, and Lesh 2001). Our innovations over these models come in the particular ways we regiment our representations and problem-solving processes to accommodate linguistic context and uncertainty. We describe these innovations further below.

The second tradition is engineering approaches to spoken dialogue systems, where researchers have shown that systems should represent the uncertainty of their automatic speech recognition results and take that uncertainty into account in their dialogue management strategies, both to accumulate information across extended interactions (Bohus and Rudnicky 2006) and to make better choices about when and how to clarify (Roy, Pineau, and Thrun 2000; Horvitz and Paek 2001; Williams and Young 2007). Such research demonstrates the simplicity and power of conceptualizing problematic interactions in conversation in terms of transient dynamics of uncertainty about what is going on. We carry this insight over into our work also.

We do not, however, adopt the formal setting for an increasing range of research on probabilistic dialogue management: the framework of partially-observable Markov decision processes (POMDPs), which casts choice as the decision-theoretic problem of

maximizing expected future utility in a stochastic environment with hidden state. This is a powerful and a flexible framework. Researchers can choose how to represent the state and action space in a POMDP model of dialogue, so it is possible to develop state models that include aspects of the hierarchical organization of coherent activity (Lemon et al. 2006; Heeman 2007; Henderson, Lemon, and Georgila 2008) as well as aspects of an evolving linguistic context (Henderson, Lemon, and Georgila 2008), and it is possible to develop action models that formalize aspects of choice in language generation (Rieser and Lemon 2008, 2009; Janarthnam and Lemon 2009). Such efforts show that POMDP models do in principle have the resources to address the issues in representation and reasoning that we are concerned with. Unfortunately, as their state and action spaces increase in size, POMDP models require increasingly formidable engineering to deal with the sparseness of data from which models are built and the computational complexity of determining optimal policies; accordingly, most work to date has focused on modeling user state, not the evolving collaboration or linguistic context in dialogue. In practice, the fine-grained representations of context and the generative action space we use in COREF would be too complex for current techniques.

We have other reasons to eschew POMDP models in our work here. In particular, we envision that dialogue systems will continue to be built with many different kinds of reasoning. On the one hand, fielded applications require consistent and intelligible decision making—a goal uniquely suited to handcrafted interaction strategies, as Paek and Pieraccini (2008), for example, highlight. On the other hand, the fundamental nature of conversation as a coordination problem suggests the possibility of using game-theoretic reasoning to make equilibrium choices in language use (Jaeger 2008), or to adopt approximations to game-theoretic reasoning such as cognitive hierarchy models (Camerer 2003), a generalization of the familiar minimax heuristic for game playing. These alternatives potentially involve quite different foundational assumptions, optimization concepts and learning algorithms from POMDPs. We take a satisficing approach to collaborative decision making (Simon 1978): we look to characterize moves in dialogue that make sense as collaborative and coherent responses to uncertainty, and focus on the knowledge the system must maintain to recognize them. We believe this methodological perspective leaves our arguments compatible with the widest possible range of frameworks for decision making in dialogue.

The third tradition we draw on is deep approaches to dialogue coherence, where researchers provide detailed models of evolving utterance context in dialogue and of the linguistic constructions that exploit this context. The strength of these approaches is their ability to account for the specific utterances available in context for speakers to signal what they have understood and where they need clarification. Deep coherence approaches adopt the perspective that each utterance in dialogue must be tightly linked to what has come before, in part because the speech act it achieves must stand in an appropriate rhetorical relation to prior discourse (Asher and Lascarides 2003) and in part because the linguistic constructions from which it is composed require specific salient material to be recovered from the dialogue context (Webber et al. 2003). Formally, the evolving context is modeled as a knowledge base, or **information state** (Bunt 2000; Larsson and Traum 2000), recording the facts needed to resolve underspecified utterance meanings. For example, anaphoric reference is mediated by a specification of which entities are available and prominent, discourse deixis is mediated by a specification of what utterances have been used with what meanings, and reasoning about coherence is mediated by a specification of the **questions under discussion** which remain to be answered in the discourse (Larsson and Traum 2000; Ginzburg and Cooper 2004). We

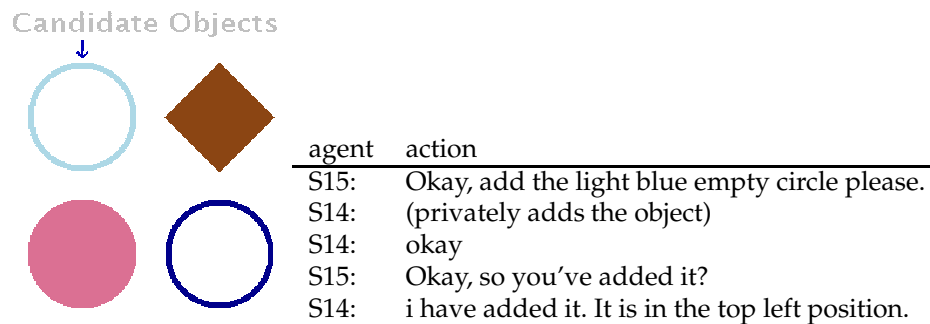


Figure 2

An ambiguous grounding action by subject S14 in a human-human dialogue.

embrace this general perspective, as well as many of the specific formalizations from prior research.

However, deep coherence models often create explanatory tension by running together descriptions of how utterances update the context with descriptions of how interlocutors manage uncertainty. Deep coherence models typically handle problematic dialogues with an **incremental common ground** approach. The idea is to distinguish public information, which can be assumed to be mutually known among interlocutors, from pending or problematic information, which remains uncertain. Rules governing speech acts with grounding functions, such as acknowledgments and clarifications, establish a formal protocol through which pending or problematic information is eventually either registered as public information or discarded as irreparable (Traum 1994; Poesio and Traum 1997; Matheson, Poesio, and Traum 2000). For example, when a new utterance occurs, its content may be marked **pending** and only recorded as public information upon a subsequent acknowledgment act by its addressee. In many cases, such rules are empirically suspect and brittle; in others they duplicate collaborative reasoning needed elsewhere in a dialogue system. Let us consider each of these problems in more detail.

To start, acknowledgments in dialogue are not in fact unambiguous signals of information becoming public (Clark and Schaefer 1989). For example, in an annotation reliability study using dialogues from the TRAINS corpus, Core and Allen (1997) found that annotators often had a hard time agreeing on whether an utterance was an acknowledgment or an acceptance. In fact, Figure 2 provides a naturally occurring fragment of human-human dialogue in COREF's domain, where interlocutors themselves treat an utterance of *okay* as ambiguous. In this interaction, S15 and S14 converse via teletype from separate rooms. S15 begins by instructing S14 to click on a certain object in S14's display. S14 does so, but S15 cannot observe the action. This leads S15 to perceive an ambiguity when S14 says *okay*: has S14 merely grounded S15's instruction, or has S14 also clicked the object? The ambiguity matters for this task, so S15 engages the ambiguity with a followup question. Such examples argue for modeling acknowledgments as giving probabilistic evidence of what has been understood and agreed in the conversation, but not as marking a transition in a formal grounding protocol.

Meanwhile, as this interaction also shows, collaborators already face a general problem of recognizing and responding to perceived ambiguities in their joint activities. The ambiguity here, about whether the object has been moved, arises generally from the fact that collaborators here have only partial information about the actions that have been taken and the contributions those actions are intended to make to the ongoing

task. Likewise, the question *you've added it?* instantiates a general strategy of asking a question to settle a significant uncertainty. Thus, although ambiguous utterances can be given a specific treatment using rules that update the discourse context in distinctive ways—for example, marking them as suspended until a special process of clarification resolves the relevant ambiguity (Ginzberg and Cooper 2004; Purver 2004)—such treatments seem couched at the wrong level of abstraction. An agent that can ask questions to resolve task ambiguities will already be able to resolve communicative ambiguities the same way.

Finally, connecting context update to the resolution of perceived ambiguities may guarantee common ground, but leaving ambiguities open can make a collaborative agent more flexible. An agent that demands a clear context but lacks the resources to clarify something may have no recourse but to take a “downdate” action—to signal to the user that their intended contribution was not understood, and discard any alternative possible contents. If the agent can proceed, however, the agent may get evidence from what happens next to resolve its uncertainty and complete the task.

We use models of collaborative activity to reconcile and bridge the insights of deep coherence approaches to linguistic context and probabilistic approaches to uncertainty management. The collaborative perspective brings two signal advantages. First, it offers an expansive understanding of the actions available to interlocutors in dialogue, and particularly flexible ways to represent their possible interrelationships and effects. We give particular attention to the task-specific collaborative problem-solving subtasks and moves involved in coordinating problematic referential communication, building on a range of previous models, particularly (Heeman and Hirst 1995; Lochbaum 1998; Blaylock, Allen, and Ferguson 2002). Moreover, as usual in collaborative models (Rich, Sidner, and Lesh 2001; Allen et al. 2007), our treatment applies uniformly not only to linguistic actions but also object manipulation and other task actions in our interface. This flexibility allows us to respect the general insights of deep coherence approaches to dialogue—particularly the use of a precise dynamic semantics to characterize the evolving utterance context and its effects on utterance interpretation—while tying ambiguity and its resolution tightly to the specific goals and expectations that prevail in our communicative setting. In particular, we use a **knowledge interface** to mediate between domain-general meanings and the domain-specific ontology required for a particular activity and setting (DeVault, Rich, and Sidner 2004). This allows us to build interpretations using domain-specific representations for referents, for task moves, and for the domain properties that characterize referents.

Second, collaborative models adopt an inherently inferential account of the link between utterances and task moves. Interpreting an utterance as an action involves recognizing the speaker’s intention: figuring out what the speaker meant to do, reconstructing implicit information and resolving ambiguity, by applying background assumptions that ultimately describe the speaker’s mental state (Perrault, Allen, and Cohen 1978; Pollack 1990). Intention recognition thus provides a framework to reconcile the complex context dependence of linguistic utterances with the fact of uncertainty in dialogue and the corresponding probabilistic representation of the state of the conversation. Moreover, intentions have a privileged status in joint activity because of the mutual commitment all collaborators have to a successful outcome (Cohen and Levesque 1990; Grosz and Kraus 1996). Computationally, this commitment results in agents organizing their choices so that their intentions are recognizable, and assuming that others do the same (Carberry 2001). The distinctive commitments and reasoning associated with contributions to collaborative activity plays out in the distinctive problem-solving mechanisms of our treatment of grounding and clarification.

Our approach builds particularly closely on our earlier work on formulating communicative intentions in dialogue (Stone et al. 2003; Stone 2004b). This work offers a particularly lightweight and flexible understanding of collaboration, in two respects. First, we use intention representations that formalize just the content of speaker's plans, and not, as in previous models (Perrault, Allen, and Cohen 1978; Pollack 1990), all the attitudes to which that content might implicitly commit the speaker. Second, we assume that collaboration emerges from specific patterns of problem solving in deliberation—particularly orchestrating actions to have recognizable intentions and interpreting actions accordingly—rather than from particular dynamics of mutual attitudes. Our lightweight, flexible approach makes our representations simpler than antecedents like Heeman and Hirst (1995) while supporting a broader range of utterance types. For example, their approach to the dialogue of Figure 1 would have interlocutors coordinating on goals and beliefs about a syntactic representation for *the dark brown object*; for us, this description and the interlocutors' commitment to it are abstract results of the underlying collaborative activity. We discuss the theoretical underpinnings of our new view and their consequences in more detail elsewhere (Stone 2004a; DeVault 2008).

Our main contribution here to the literature on collaboration in dialogue, then, consists in the arguments we present that uncertainty needs to be put front and center in tracking contributions to conversation. Previous collaborative systems offer only limited abilities to cope with ambiguity and incomplete information (Lesh, Rich, and Sidner 2001; Allen et al. 2007). Our preliminary implementations of collaborative reference (DeVault et al. 2005) also avoided uncertainty about the context. Initially, in fact, we saw it as a theoretical challenge just to reconcile the idea of uncertainty in context with established pragmatic theories (DeVault and Stone 2006; Thomason, Stone, and DeVault 2006), in light of Stalnaker's influential identification of conversational context with interlocutors' common knowledge, an inherently bivalent construct (Stalnaker 1974). Our short paper (2007) offers a brief overview of our model, implementation and experiment, but the results and arguments that are most important in this article had yet to take shape. Some can be found in preliminary form in DeVault's unpublished PhD dissertation (2008), along with a complete description of COREF and its implementation.

3. Contribution Tracking

We present our ideas through a detailed analysis of a referential communication task studied in pairs of human subjects by Clark and Wilkes-Gibbs (1986). Each interlocutor perceives a collection of visual objects, as illustrated in Figures 1–2. The interlocutors perceive identical objects, but with shuffled spatial locations. One interlocutor, who we call the director, sees a target object highlighted on their display with an arrow, and is charged with conveying to their partner, who we call the matcher, which of the displayed objects is the target. The interlocutors go through the objects one by one, with the matcher attempting to identify and click on the correct target at each step.

In Section 3.1, we show how we regiment collaboration, communicative action and uncertainty in this domain. We characterize collaboration partly in terms of an evolving task state and partly in terms of patterns of possible coherent activity. The task state registers the progress interlocutors make in negotiating problem-solving efforts and in achieving domain goals. In referential communication, for example, when interlocutors work to build a description that uniquely characterizes a target object, the task state records the descriptive attributes that have been contributed and the set of alternative objects that share those attributes and remain possible candidates for reference. The task state also manages linguistic context dependence, by recording such things as

the utterances that have been used and the discourse referents that have been made salient. Patterns of task activity organize the collaboration into a hierarchy of ongoing subtasks, and directly constrain the moves that interlocutors can coherently make, either implicitly or explicitly, in each state. By drawing on these background constraints, we can formalize the problem of recognizing the speaker's intention in using an utterance as a problem of abductive inference: we must interpret the logical form of the utterance in context as a signal of the coherent sequence of task actions the speaker is committed to carry out next (Hobbs et al. 1993).

In Section 3.2, we describe the process of **contribution tracking** as a set of reasoning capabilities defined over our dialogue representations. We put particular focus on the generative mechanisms that lead our system, COREF, to produce clarification requests like the example *do you mean dark brown* of Figure 1. To start, COREF's reasoning exposes ambiguity about what the user means as uncertainty in the dialogue state that results from the user's utterance. Here COREF assumes that the user intends to identify the color of the target object with *it is brown* and therefore finds two possible interpretations: one for the dark brown color of the empty diamond and one for the light brown color of the solid square. After the utterance, COREF is uncertain about which meaning was intended and thus which constraint the user has contributed.

Second, COREF's dialogue strategies are formulated for an uncertain dialogue state. This allows COREF to proceed with appropriate high-level dialogue moves despite having more than one alternative for what the context is. Here COREF settles on a clarification move, because we have specified a policy of that COREF should clarify when its alternatives describe different constraints on the target object. For other kinds of uncertainty, COREF might proceed without clarifying.

Third, COREF's reasoning in generation aims to synthesize utterances for which the user will recover a specific and useful interpretation no matter what the context is. Here COREF explicitly constructs the utterance *do you mean dark brown* by carrying out an incremental derivation using a lexicalized grammar. The rich representation of the utterance context allows the system to recognize the applicability of forms that cohere with what has gone before, such as the use of the frame *do you mean* to refer to content from the previous utterance, whatever it may have been. The model predicts that this underspecification is unproblematic, but predicts that the ambiguity of *brown* must be eliminated and therefore motivates the adjunction of the modifier *dark*.

3.1 Actions, Tasks, State and Uncertainty

The context for COREF describes both the state of the ongoing referential activity and the semantic and pragmatic status of information in the dialogue. Foundationally, we understand this context as a product of prior interlocutor action, one ultimately determined by the understood conventions through which people normally coordinate referential communication (Lewis 1979; DeVault and Stone 2006). We therefore assume a quite indirect relationship between the collaborative context and constructs like mutual belief that describe interlocutors' occurrent mental states.

The fundamental unit of description for context dynamics is the **task action**, an abstract move that effects a contribution to the ongoing activity. Task actions can be closely associated with overt behaviors carried out in public by interlocutors, or they may be tacit, so that some interlocutors have no direct evidence that the action has taken place (Thomason, Stone, and DeVault 2006). Overt actions in COREF's domain include interface actions that update both interlocutors' displays, as well as the actions that are grammatically associated with linguistic material. Such actions include both

the main point of an utterance, for example asking a question, as well as incidental, grammatically-specified changes to the context, for example raising a discourse referent to salience. Tacit actions in COREF's domain include interface actions that update only one interlocutor's display, such as moving objects around the display, as well as task-relevant cognitive actions, such as identifying the target object or implicitly initiating, completing, or abandoning a subtask. A speaker is free to use tacit actions as well as overt task actions to update the context. However, successful coordination requires the speaker to provide sufficient evidence in their overt behavior to reconstruct any tacit actions they have committed to.

Task actions are organized into broader patterns of coherent activity through a set of task networks describing the possible sequences of actions that interlocutors may use coherently to complete a task. For example, COREF has a single-object reference task that tracks how a director D and matcher M together set up and solve a constraint-satisfaction problem to identify a target object. In any state, D and M have agreed on a target variable T and a set of constraints that the value of T must satisfy. When M recognizes that these constraints identify R , the task ends successfully. Until then, D can take actions that contribute new constraints on R . Since what D says adds to what is already known about R , the identification of R can be accomplished across multiple sentences with heterogeneous syntactic structure. The inventory of tasks in COREF also includes an overall multi-object reference task, a yes/no question task, a reminder question task, a clarification task, and an ambiguity-management task that is automatically pushed after each utterance or action. In addition to task networks, COREF includes a small set of rules which allow an additional range of general-purpose actions to occur at a wide variety of points in the dialogue. These include asking a yes/no or wh-question, describing what has just happened in the interaction, and abandoning a task currently under way. DeVault (2008) gives full details about the specific task networks and general rules implemented in COREF.

The dialogue state, meanwhile, offers a representation of the instantaneous status of the interaction. We specify not only a stack of tasks that are underway but also use appropriate task-specific data structures to record the progress that has been made in each. For example, for each collaborative reference task, the dialogue state includes a constraint network; the constraint network represents the candidate objects that might be the target, the properties the target must have, given the assertions that have been made so far, and the properties that the target cannot have, given the assertions that have been denied so far. This allows the context to track progress towards identifying the target uniquely. Analogously, the dialogue context includes representations of the extralinguistic context, including the objects and properties visible in the display, and representations of the linguistic context, including aspects of the discourse history such as specifications of recent utterances and of salient referents.

A primitive update operation $\text{do}(c, a)$ describes the new dialogue state that results after the task action a is executed in dialogue state c . By iterating this update operation, we can describe the composite effect $\text{do}^*(c, A)$ of a sequence of task actions A . This in turn determines what happens when interlocutors take overt behaviors. Interlocutors contribute to conversation by manifesting a suitable communicative intention—in other words by carrying out an observable behavior with a specific commitment as to how that behavior will generate a sequence of actions that links up with and updates the dialogue state. We formalize the effects of an intention i using behavior b to generate a sequence of task actions A through a function $\text{update}(c, i)$. The same model of update is used for both interlocutors—director and matcher, user and system. This function first revises the state to record the fact that behavior b was observed and signaled intention

i as part of the history of the interaction; then the function updates the state with the effects of the action sequence A ; and finally it revises the ongoing task by pushing an ambiguity management subtask which affords interlocutors the chance to acknowledge and clarify how the behavior was to be interpreted in subsequent discourse, before proceeding further with the task. In fact, the internal representations of intentions in COREF actually formalize the **explanation** why the speaker would expect their observable behavior in context to generate the actions they intend (Stone 2004a). However, for ease of exposition in this paper, we will usually indicate the actions that an interlocutor commits to as a shorthand for their intention.

Of course, interlocutors actually produce and observe behaviors rather than actions. They must recognize intentions by inference. We assume that the speaker applies the same problem-solving characterization in formulating utterances and intentions as the addressee applies in recognizing intentions from utterances, and COREF uses the same model both in understanding and in generation and as director and as matcher. In particular, we regiment calculating the interpretation of an utterance as an abductive constraint satisfaction problem. One source of constraints is the task network that describes possible actions in context. The other is background knowledge about the kinds of intentions associated with specific observable behaviors. This is particularly important for utterances, where the interpretive background is specified by the grammar. In particular, the grammar associates utterances with schematic specifications of task moves that contain variables. These variables are subject to grammatically-specified presupposed constraints. These constraints must be satisfied in the current dialogue state on the intended resolution of the variables. An abductive process of constraint satisfaction is necessary because the interpreter must be prepared to hypothesize that the current dialogue state has been implicitly updated by tacit actions which are not yet associated with observable behaviors.

In this constraint satisfaction process, the interpreter's evidence may still leave multiple options open. When it does, the interpreter is uncertain about the dialogue state that results from the action.

We will use the example subdialogue in Figure 1 to illustrate how these representations describe problematic interactions in ways that allow interlocutors to recognize, react to, and overcome their transient uncertainty. Figure 3 depicts the evolving representations of context in this dialogue as maintained utterance-by-utterance by COREF. The topmost row, under the heading 'EI' (for 'Experiment Interface'), shows COREF's perspective on the visual objects. This perspective grounds out in a database describing these objects and their properties as part of the dialogue state. An ellipsis indicates that COREF's version of the experiment interface has not changed from one step to the next, so that the corresponding database carries over unchanged.

The second row indicates the "time". Time ticks with each observable event that moves the interaction forward. The observable event ('OE') that is next to occur is indicated in the third row.

The fourth row schematizes the intentions that COREF attributes to interlocutors to move the conversation forward, and the updates that those intentions trigger. It illustrates the evolving dialogue states – with state identifiers like $s55812$, $s55979$, etc. – and shows how they are interrelated by the interpretations that COREF assigns to observed events. It describes each interpretation in terms of the tacit actions that COREF hypothesizes to make the utterance coherent, and the overt actions that COREF takes the speaker to be committed to, on that interpretation. The uncertainty COREF faces is indicated by the presence of alternative states. The presence of a single node, such as $s55812$ in the figure, indicates that COREF is completely certain. By contrast, the

alternatives present at times 4, and 5 shows that COREF is temporarily uncertain which of two dialogue states is the correct one.

Figure 3: COREF eliminates a perceived ambiguity

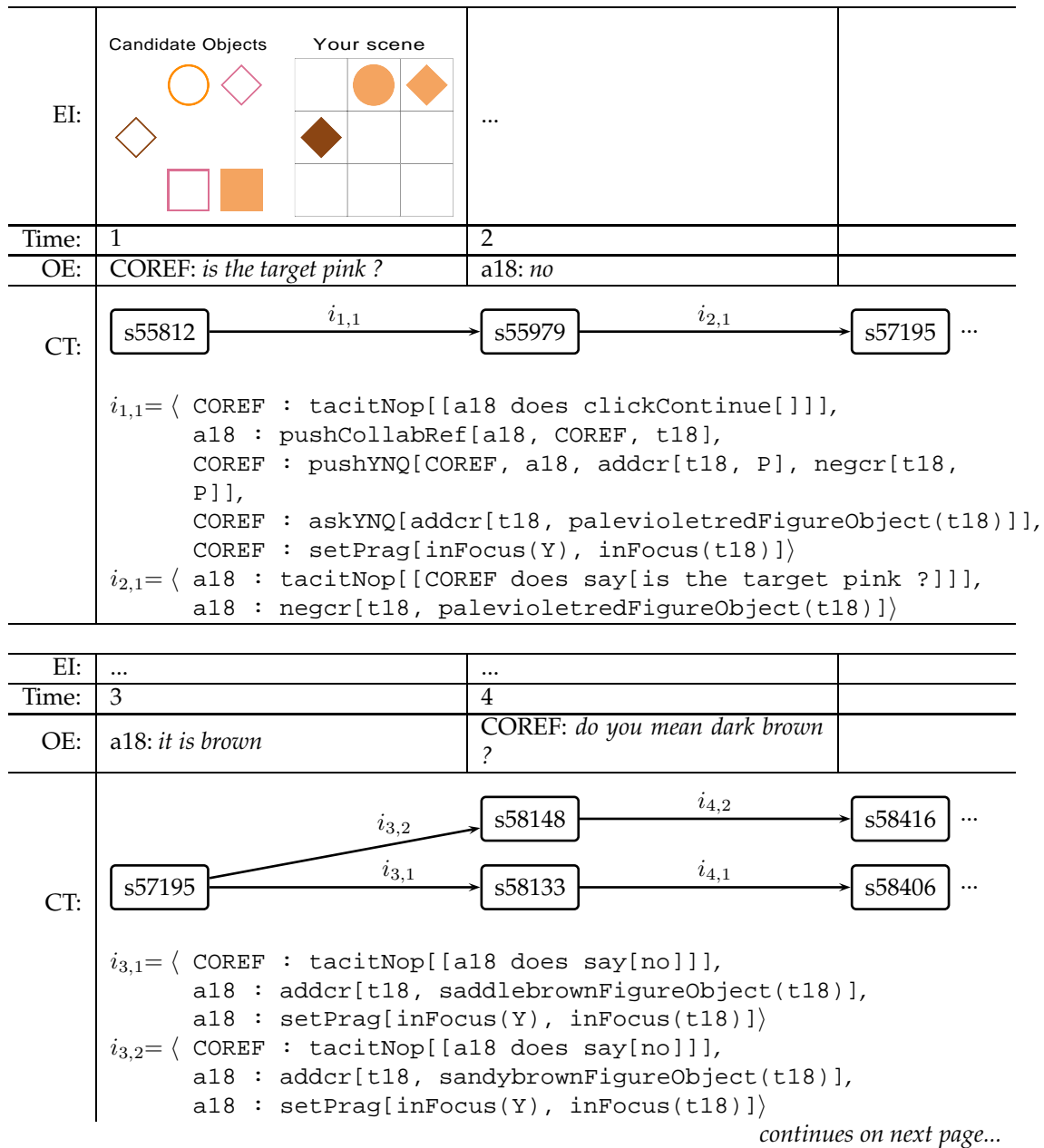


Figure 3: COREF eliminates a perceived ambiguity (continued)

$i_{4,1} = \langle$ COREF : pushClarify[[a18 does say[it is brown]], a18, COREF, Brown], COREF : pushYNQ[COREF, a18, addcr[Brown, P], negcr[Brown, P]], COREF : askYNQ[addcr[Brown, equal(Brown, saddlebrownFigureObject)]] $i_{4,2} = \langle$ COREF : pushClarify[[a18 does say[it is brown]], a18, COREF, Brown], COREF : pushYNQ[COREF, a18, addcr[Brown, P], negcr[Brown, P]], COREF : askYNQ[addcr[Brown, equal(Brown, saddlebrownFigureObject)]] 									
<table border="1" style="width: 100%; border-collapse: collapse;"> <tr> <td style="width: 33%;">EI: ...</td> <td style="width: 33%;">...</td> <td style="width: 33%;"></td> </tr> <tr> <td>Time: 5</td> <td>6</td> <td></td> </tr> <tr> <td>OE: a18: <i>yes</i></td> <td>COREF: <i>done</i></td> <td></td> </tr> </table>	EI:		Time: 5	6		OE: a18: <i>yes</i>	COREF: <i>done</i>	
EI:								
Time: 5	6								
OE: a18: <i>yes</i>	COREF: <i>done</i>								
<p>CT:</p> <div style="margin-left: 20px;"> <div style="border: 1px solid black; padding: 2px; display: inline-block; margin-right: 10px;">s58416</div> X </div> <div style="margin-left: 20px;"> <div style="border: 1px solid black; padding: 2px; display: inline-block; margin-right: 10px;">s58406</div> $\xrightarrow{i_{5,1}}$ <div style="border: 1px solid black; padding: 2px; display: inline-block; margin-right: 10px;">s60204</div> $\xrightarrow{i_{6,1}}$ <div style="border: 1px solid black; padding: 2px; display: inline-block; margin-right: 10px;">s60320</div> ... </div> $i_{5,1} = \langle$ a18 : tacitNop[[COREF does say[do you mean dark brown ?]]], a18 : addcr[Brown, equal(Brown, saddlebrownFigureObject)] $i_{6,1} = \langle$ COREF : tacitNop[[a18 does say[yes]]], COREF : setVarValue[Brown, saddlebrownFigureObject], COREF : setVarValue[t18, rh656899_4], COREF : addToScene[rh656899_4], COREF : past[COREF, addToScene[rh656899_4]] 									

Inspecting the intention representations more closely allows us to make our discussion thus far more concrete. We begin with COREF's utterance, *is the target pink?* for which the intention is schematized as $i_{1,1}$. At this point in the dialogue, Subject A18 has clicked through to reveal the next object they had to describe as the director. Accordingly, intention $i_{1,1}$ begins by hypothesizing a sequence of three tacit actions that set the stage for COREF's question within the overall activity: first, the move where COREF implicitly passes (*tacitNop*) on the opportunity to clarify the meaning of its interlocutor's last interface action; second, the move where the interlocutors tacitly begin to collaborate on a new collaborative reference task (*pushCollabRef*) on the next target, $t18$; and third, the move where COREF tacitly begins a nested segment of the discourse (*pushYNQ*) aimed at adding or rejecting a constraint about the target. The intention concludes with the two actions that are associated overtly with the utterance: a move actually asking the question whether the target is pink (in domain terms, whether it has the property *palevioletredFigureObject*); and a grammatically-specified move replacing whatever previously had the pragmatic status *inFocus* with the discourse referent $t18$ for the target.

The reply *no* gets associated with the intention $i_{2,1}$. As before, it begins by implicitly passing on the opportunity to clarify the meaning of COREF's utterance. It goes on (with `negcr`) overtly to reject the constraint COREF has proposed.

The user continues with *it is brown*. The ambiguity leads to two possible intentions, represented as $i_{3,1}$ and $i_{3,2}$. The two intentions agree in structure in this case. Both begin by assuming that COREF implicitly passes on the opportunity to clarify on the answer the user has just provided; this pops not only the ambiguity management subtask but also the question subtask COREF has initiated, so that we return to the overall collaborative reference process to identify t_{18} . To this task, the user now overtly adds a constraint (`addcr`). It is either, for $i_{3,1}$, that the target has the dark brown color represented as `saddlebrownFigureObject`, or, for $i_{3,2}$, that the target has the light brown color represented as `sandybrownFigureObject`. Finally, as before, the utterance also effects the overt move of maintaining the discourse referent t_{18} with the status `inFocus`. The ambiguity in interpretation at step 3 is exposed as ambiguity in about the conversational state at step 4.

The response given by COREF at step 4 takes the opportunity to make a series of moves within the ambiguity management task introduced with the user's utterance at step 3. Because the interpretation of an utterance is inherently defined only relative to a specific context, we represent the interpretation that the utterance gets in state s_{58133} , as represented in $i_{4,1}$, separately from the interpretation that it gets in state s_{58148} , as represented in $i_{4,2}$. In fact, though, in this case, those two intentions have the same content. This content is to begin a clarification subdialogue, which has the same logical structure as any other dialogue for collaborative reference, except that the target entity that the interlocutors are working to identify is in fact a specific referent involved in the meaning of the last utterance. In this case, the target variable `Brown` represents whatever the user meant by *brown* in utterance 3. The move that COREF makes in this subdialogue is exactly parallel to the move it made in utterance 1: COREF pushes a subdialogue to determine whether the target value is actually identical to the domain property `saddlebrownFigureObject`, and raises the question with the move `askYNQ` which is overtly associated with COREF's question.

At move 5 comes the reply *yes*. In state s_{58406} this reply is interpreted as $i_{5,1}$: the speaker passes up the chance to address ambiguity in COREF's clarification task, and goes on to add the constraint that the variable `Brown` should in fact equal the domain property `saddlebrownFigureObject`. In state s_{58416} , however, this utterance has no coherent interpretation. The anaphoric structure of the speaker's answer, in this discourse context, determines that the utterance could only serve as to add the constraint that `Brown` should be `saddlebrownFigureObject`. Meanwhile, this state already records the fact that `Brown` must be `sandybrownFigureObject`, because that's what the user meant in $i_{3,2}$. There is no way to reconcile these incompatible constraints.

The dialogue continues in the unambiguous state s_{60204} . This state reflects the resolved intended meaning for the user's utterance with *it is brown*; the corresponding constraint is enough to identify the object that must be the target for this interaction. Thus, COREF now completes the interaction, and offers an utterance that allows its inferences and actions to be recognized. The intention at $i_{6,1}$ shows that COREF passes on the chance to clarify *yes*, resolves the clarification subdialogue (with the first `setVarValue` action), and goes on to resolve the overall collaborative reference subtask (with the second `setVarValue`), and then tacitly indicates the choice it has made in the interface (with the action `addToScene`). The only overt action, directly associated with *done*, is the last one, an assertion (`past`) which adds to the record of the interaction the

proposition that the salient task-relevant move was made in the task—in this case, adding the object.

3.2 Reasoning

The representations of Section 3.1 allow us to describe the problem solving involved in contribution tracking formally. We also offer concrete examples of this reasoning through the example of Figure 3. An important theme is that, throughout such interactions, COREF faces considerable uncertainty about all of the following: what contributions its human interlocutor is making with their utterances, what dialogue state they are in, and even potentially what contributions COREF itself is making with *its* utterances. Despite this uncertainty, COREF is able to plan novel, coherent and recognizable utterances, and often achieve successful task outcomes.

3.2.1 Filtering. The basic interpretive operation in contribution tracking is not *updating*—that is, tracking unambiguous changes to the dialogue state—but *filtering*—propagating uncertainty about the dialogue state at time t to uncertainty about the dialogue state at time $t + 1$ based on an observed behavior.

To characterize filtering in the setting of collaborative activity, we need to describe more precisely how tacit actions figure in the intention recognition process. Formally, for any state c and interlocutor S , we can use the next actions that could contribute to the pending tasks in c to determine a set of alternative states $Z(c, S)$ that could be reached by S from c just using tacit actions. We call this set of alternative states the *horizon*. See Thomason, Stone, and DeVault (2006). Let us write $c : i$ to denote an interpretation which shows the speaker (or actor) acting in state c with a commitment to intention i . In understanding, an agent H starts from a prior probability distribution over the initial context at time t given the evidence E_t available so far: $P_H(c_t | E_t)$. H observes an behavior b_t (carried out by agent S), and must infer $\hat{c}_t : i_t$ to explain it. H can assume that the new state \hat{c}_t must be some element of $Z(c_t, S)$, and that i_t must match behavior b_t into \hat{c}_t so as to contribute to the ongoing tasks. H will inevitably bring substantial background knowledge to bear, such as grammatical knowledge and interpretive preferences. However, H 's evidence may still leave multiple options open. We summarize H 's intention recognition as a probabilistic likelihood model $P_H(\hat{c}_t : i_t | c_t, b_t)$. (As usual, we assume the current state tells you everything that could in principle influence the interpretation of the current action.) Filtering combines update, prior and likelihood:

$$P_H(c_{t+1} | E_{t+1}) = P_H(c_{t+1} | b_t, E_t) \propto \sum P_H(\hat{c}_t : i_t | c_t, b_t) P_H(c_t | E_t)$$

where the summation ranges over all values of c_t , \hat{c}_t , and i_t such that $c_{t+1} = \text{update}(\hat{c}_t, i_t)$.

This definition of filtering accounts for the reasoning needed to track the evolving dialogue state as depicted in Figure 3. As an illustration, let's focus on the third utterance, *it is brown*. We start with just one possible state, $s57195$, which must of course be assigned probability one. This corresponds to the statement of the prior in our filtering operation: $P(c_3 = s57195 | E_3) = 1$. Now we find two possible interpretations: the intentions given as $i_{3,1}$ and $i_{3,2}$ in Figure 3, corresponding to the two different colors that might be meant by *brown*. As noted earlier, each of these intentions involves a tacit move to a new context not explicitly represented in the figure—call it $s57196$ —which implicitly completes any discussion of the contribution of the user's previous

utterance *no*. Thus in this case we have $s57196 \in Z(s57195, S)$. In fact, COREF’s intention recognition model assigns a specific probability to the choice of this tacit move, which may differ from other possibilities on the horizon. However, since this move is shared between all the different interpretations COREF finds for this utterance, this factor is normalized away. Meanwhile, COREF’s constraint satisfaction model happens to assign equal probability to the two resolutions for the meaning of *brown* in state $s57196$. Thus the relevant likelihood terms for this operation are:

$$\begin{aligned} P_H(\hat{c}_3 = s57916 : i_3 = i_{3,1} | c_3 = s57195, b_3 = \textit{it is brown}) &= 0.5 \\ P_H(\hat{c}_3 = s57916 : i_3 = i_{3,2} | c_3 = s57195, b_3 = \textit{it is brown}) &= 0.5 \end{aligned}$$

As described in the figure, $\text{update}(s57196, i_{3,1}) = s58133$ and $\text{update}(s57196, i_{3,2}) = s58148$. Thus the filtering operation combines COREF’s prior and likelihood models to track the new state as expected:

$$\begin{aligned} P_H(c_4 = s58133 | E_4) &= 0.5 \\ P_H(c_4 = s58148 | E_4) &= 0.5 \end{aligned}$$

3.2.2 Collaborative response. Effective dialogue management under uncertainty requires agents to collaborate actively to resolve ambiguities with their interlocutors. We regiment these collaborative skills in contribution tracking by stipulating that agents have an operation that specifies which moves would be **acceptable** given an agent’s current uncertainty about the dialogue state.

In COREF, this operation is realized as a hand-built policy. This policy includes COREF’s knowledge about how to move the task forward constructively. For example, as director, COREF is only willing to assert constraints that it knows are true of the target object. And COREF is never willing to trigger the failsafe button on the interface that skips the current object and moves on to the next one. The policy also specifies how COREF should deal with uncertainty. For example, COREF’s policy deems it acceptable to ask for clarification any time COREF is uncertain which constraint a speaker intended to add with an utterance, as in Figure 1. Similarly, COREF’s action policy deems it acceptable for the agent to ask whether a non-public action m has occurred, if some possible dialogue states but not others indicate that m has taken place. For example, COREF translates an ambiguous acknowledgment like that of Figure 2 into uncertainty about whether the “add object” action has tacitly occurred in the actual dialogue state; COREF follows up such an *okay* by asking *did you add it?*

3.2.3 Minimizing ambiguity. The other key skill for contribution tracking is to generate natural language utterances that do not exacerbate the problems of ambiguity even when used in uncertain contexts. This way agents can continue to make contributions their interlocutors understand, even in cases of transient uncertainty. We connect this ability to agents’ reasoning by requiring the generation module in the system to produce utterances whose interpretations are weakly recognizable in a sense we now specify.

We assume that agents can take their own probabilistic models of interpretation as good indicators of their partners’ disambiguation preferences, and can therefore discard interpretations whose probability falls below a threshold ϵ . They are of sufficiently low probability, relative to others, that they can safely be ignored. This formalizes the idea that collaborators are working actively to coordinate their actions and behavior and to disambiguate their contributions to their joint activity.

Consider then an observable action b by S . If there were a unique dialogue state c established with certainty among the interlocutors, then the set of recognized interpretations for b would be $R(c, b) = \{\hat{c}: i | P(\hat{c}: i | c, b) \gg \epsilon\}$. However, in general, S is uncertain which of $C = \{c_1, \dots, c_k\}$ is the actual dialogue state, and expects that H may give any of these a high prior and take seriously the corresponding interpretations of the utterances. Indeed, S must also be prepared that S is actually making any of these contributions! Thus, H and S should consider any interpretation in $R^*(C, b) = \cup_{c \in C} R(c, b)$. $R^*(C, b)$ is **weakly recognizable** if and only if each $c_i \in C$ is associated with at most one interpretation in $R^*(C, b)$.

The formalism explains why, in generation, COREF chooses to elaborate its utterance *do you mean brown?* by adding the word *dark*. COREF's policy makes a clarification question acceptable across all of the candidate contexts after the user says *it is brown*. But *do you mean brown?* is not weakly recognizable. For example, in *s58148*, there are two interpretations, which could be paraphrased *do you mean light brown?* and *do you mean dark brown?* COREF therefore chooses to coordinate more finely on the alternative interpretations of its clarification action. The utterance *do you mean dark brown?* has only one interpretation in each of *s58133* and *s58148* and therefore represents a solution to COREF's communicative problem.

In Thomason, Stone, and DeVault (2006) we investigate a stronger discipline of recognizability, where each utterance results in a **checkpoint**, where speaker and hearer agree not only on a unique interpretation for the utterance but also on a unique resulting context. Enforcing this constraint supports the traditional attribution of mutual knowledge to the two interlocutors at each point in the conversation. By contrast, the weak recognizability we enforce in contribution tracking allows for uncertain contexts and makes interpretation more robust to potential differences in interlocutors' perspectives. In interpreting a user utterance, COREF expects to find zero, one, or multiple interpretations in each possible context. In generation, COREF is sometimes willing to take the risk of using an action or utterance that may not be interpretable in all possible contexts. Taken together, this means new utterances can serve not only to present the speaker's intention, but also in some cases to introduce or defuse uncertainties about the true context. Checkpoints, where COREF achieves certainty about the true context, arise as side effects of its collaborative activity and its inferential processes, rather than as a strict requirement in the architecture. While there is no guarantee that any given speaker contribution will ever become common ground, COREF's dialogue policies are designed to try to achieve common ground when it is practical to do so.

4. Implementation and data collection

We implemented COREF in Java. A set of interface types describes the flow of information and control through the architecture. The representation and reasoning outlined in Section 3 is accomplished by implementations of these interfaces that realize our approach. Modules in the architecture exchange messages about events and their interpretations. (1) Deliberation responds to changes in the dialogue by proposing task actions. (2) Generation constructs collaborative intentions to accomplish the planned task actions. (3) Understanding infers collaborative intentions behind user behaviors. Generation and understanding share code to construct intentions for utterances, and both carry out a form of inference to the best explanation. (4) Update advances the dialogue state symmetrically in response to intentions signaled by the system or recognized from the user; the symmetric architecture frees the designer from programming complementary updates in a symmetrical way. Additional supporting infrastructure

handles the recognition of input actions, the realization of output actions, and interfacing between domain knowledge and linguistic resources.

4.1 Assumptions and heuristics

The main challenge in COREF’s implementation concerns the efficient management of the search operations inherent in general problem-solving mechanisms. Our implementation therefore adopts various assumptions and heuristics in understanding, dialogue management and generation.

In understanding, when an actor A performs an observable behavior b at time t , exhaustive search would involve an attempt to solve the constraints associated with b at each state $s \in Z(s_k, A)$ such that $P_H(c_t = s_k | E_t) > 0$. This calculation is extravagant: there are many tacit actions that could be performed coherently, according to the agent’s task models, but which would be irrelevant to the particular utterance at hand. Consider *ok* for example. Because *ok* cannot be grammatically analyzed as a question, it is useless to hypothesize a wide range of tacit actions as precursors to *ok*, including raising a question or initiating clarification, because these tacit actions all require an overt questioning action to follow. Thus, COREF currently applies a set of hand-built rules to prune away specific parts of the horizon graph on the basis of the possible dialogue moves that the utterance might achieve. Additional pruning limits the total depth of the horizon graph.

To interpret an utterance, COREF analyzes it according to a hand-built lexicalized grammar. The specific grammar formalism is TAGLET (Stone 2002), a variant of tree-adjointing grammar in which the full adjunction operation is replaced with left and right sister adjunction, and in which lexical entries are linked to semantic constraints on variable values. The overall constraints associated with an utterance are determined by performing a bottom-up chart parse of the utterance, and conjoining the presuppositions and dialogue acts associated with each edge in the chart. Presupposed constraints are in fact solved incrementally in the chart, using a hand-built domain model which directly associates particular constraints with algorithms that deliver assignments to variables. In total, COREF’s grammar contains 379 hand-built lexical entries.

To assign probabilities, the COREF implementation uses a hand-built model which simply assigns probability inversely proportional to the “size” of the intention:

$$P(I = i_{t,j} | o, S_t = s_k) \propto \frac{1}{\text{NUM-TACIT-ACTIONS}(i_{t,j}) + 1}$$

This hand-built model measures the size of an intention as the number of tacit actions in the intention, plus 1 for the public action that concludes each intention. In DeVault and Stone (2009) we show how COREF’s interactions with its users can be used automatically to create a probabilistic intention-recognition model that weights fine-grained features of intentions and the dialogue states in which they are hypothesized.

To keep dialogue management tractable for real-time interaction, COREF represents its uncertainty through a maximum of three alternative dialogue states. If more than three states are possible, the three most probable are retained, and the others discarded. Further, after each object is completed (by the director taking a public `continueTask(T)` action, or either agent taking a public `skip(T)` action), COREF discards all but the most probable state, to avoid retaining unilluminating historical ambiguities. If either of these adjustments is made, the probability mass is renormalized across the remaining viable states. For robustness, COREF also implements special-case reasoning for observations that have no interpretations in its models. In this case, COREF

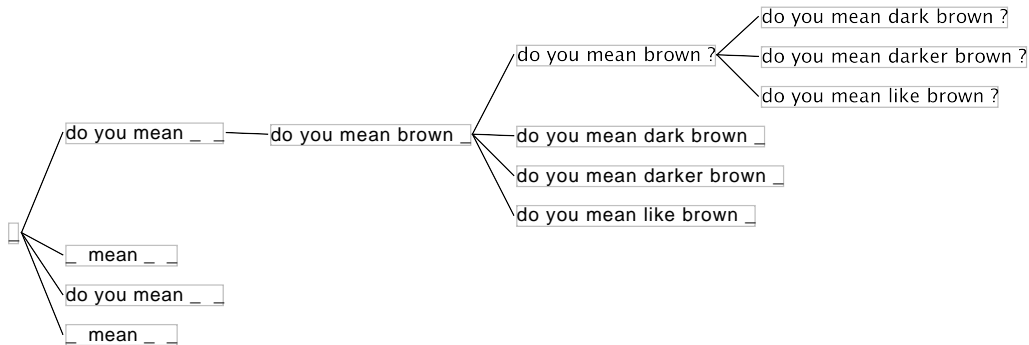


Figure 4

SPUD generates *do you mean dark brown?* for COREF. In this figure, each search node considered by SPUD is depicted using only the surface text at that search node. Nodes are ranked vertically, with higher ranked nodes above lower ranked nodes. An underscore indicates the presence of a syntactic gap in the provisional syntactic structure at the node.

translates each previously viable state $S_t = s_t$ into a new viable state by performing book-keeping and uncertainty management updates—effectively recording that the observation was made and that it was problematic, but nothing more.

For COREF’s generation process, we adapt established natural language generation techniques to a setting of task-oriented generation under uncertainty. Our basic building block is the SPUD generation algorithm of Stone et al. (2003). SPUD is a search-based generation algorithm that uses a lexicalized tree-adjoining grammar to incrementally construct an utterance that achieves a communicative goal (in COREF’s case, a dialogue act) in a way that is unambiguous in a specific context. COREF’s implementation of SPUD uses the same grammar and solver for presupposed constraints as used in understanding. Because SPUD was designed to operate under certainty about the context in which the utterance occurs, however, COREF doesn’t use SPUD directly in generation.

Before generation, COREF first constructs both the possible and acceptable horizon graphs for the dialogue states it thinks it might be in. Crucially, while the acceptable horizon graph circumscribes the communicative goals COREF is willing to adopt, it is the **possible** horizon graph that is used in anticipating user interpretations. COREF uses this horizon graph to construct an input context for SPUD that reflects the ambiguities that arise throughout the possible horizon. Once SPUD has produced a candidate utterance, COREF explores the alternative interpretations that arise for that utterance in specific states across the possible horizon. If these interpretations only contain acceptable actions, then COREF adopts the output utterance. Otherwise COREF considers making a different dialogue move instead. In the case that multiple interpretations are supported by an utterance, if COREF accepts the utterance, we view COREF as *willing to be interpreted as making any of those contributions*. We refer to this scenario as one in which COREF makes an underspecified contribution.

Consider time 4 in the object subdialogue of Figure 3, where COREF asks *do you mean dark brown?*. Figure 4 depicts SPUD’s search process for this utterance. Note in particular that SPUD considers simply generating the question *do you mean brown?* before settling on *do you mean dark brown?* It continues to refine the question, however, because its interpretation remains ambiguous. In this way, COREF avoids using a simple word like *brown* when COREF itself would perceive that word as ambiguous in the current context.

<i>correct</i>	<i>no object</i>	<i>skipped</i>	<i>wrong</i>
75.0%	14.3%	7.4%	3.3%

Table 1

Overall distribution of object outcomes.

4.2 Data collection

The result of these implementation decisions is a dialogue agent that can carry out conversations at interactive rates over a teletype interface. With this system in hand, we recruited 20 human subjects to carry out a series of collaborative reference tasks. Most of the subjects were undergraduate students participating for course credit at Rutgers University. The study was web-based; subjects participated from the location of their choice, and learned the task by reading on-screen instructions. The task instructions informed each subject that they would work with an interactive dialogue agent rather than a human partner.

Each subject worked one-by-one through a series of 29 target objects with COREF, for a total of 580 objects and 3245 utterances across all subjects. For each subject, the 29 target objects were organized into 3 groups, with the first group of 4 in a 2x2 matrix, the next 9 in a 3x3 matrix, and the final 16 in a 4x4 matrix. As each object was completed, the correct target was removed from its group, leaving one fewer object in the matrix containing the remaining targets. The roles of director and matcher alternated with each group of objects. Either COREF or the subject was randomly chosen to be director first.

The experiment interface allows an object to be completed with one of four outcomes. At any time, the matcher can click on an object to add it to their scene, which is another matrix containing previously added objects for the same group. An object is completed when the director presses either the Continue (next object) button or the Skip this object button, or when the matcher presses the Skip this object button. An outcome is scored *correct* if the director presses Continue (next object) after the matcher has added the correct target to her scene. It is scored *skipped* if the human subject presses the Skip this object button. (COREF never presses the Skip this object button.) It is scored *no object* or *wrong* if the director presses Continue (next object) before the matcher adds any object, or after the matcher adds the wrong object, respectively.

5. Results

We begin with an overview of how COREF's performance relates to its uncertainty about dialogue state. Table 1 summarizes COREF's overall performance in the task. Table 2, meanwhile, shows the distribution in the number of states perceived as viable by COREF across all subjects and dialogue events. These data show that COREF is usually completely certain what the state is—In fact, for 436 (75.2%) of the 580 object subdialogues, COREF did not face any uncertainty at any point. In the remaining 144 (24.8%) object subdialogues, COREF faced varying amounts; in all, COREF is uncertain 16.6% of the time.¹

¹ Since COREF truncates its uncertainty at 3 possible states, the higher frequency of 3 possible states relative to 2 in Table 2 masks a longer underlying tail.

1 context	2 contexts	3 contexts
83.4%	6.8%	9.8%

Table 2

Number of possible contexts perceived when utterances or actions occur.

To better understand how uncertainty affects outcome, we investigated COREF’s performance on individual objects as a function of uncertainty. To summarize the effect of this uncertainty on the agent’s performance, we computed the mean uncertainty for the subdialogue addressing each object. We defined the mean uncertainty for a subdialogue as the mean number of states that COREF viewed as viable when it interpreted the actions and utterances that occurred during the subdialogue (including actions and utterances performed both by the user and by COREF).

Figure 5 shows the agent’s mean uncertainty during object subdialogues as a function of the outcomes of those subdialogues. By pair-wise Wilcoxon rank sum tests,² the only significant differences in these distributions are between *correct* and *no object* outcomes ($W = 14309.5, p < 0.0001$), and between *correct* and *skipped* outcomes ($W = 7594, p < 0.005$). These differences show that the agent’s success at achieving a *correct* outcome, rather than a *skipped* or *no object* outcome, is related to the agent’s mean uncertainty during the object subdialogue.

Figure 6 offers a complementary perspective, showing object outcomes as a function of the agent’s mean uncertainty during object subdialogues. The figure shows that COREF’s performance, when measured by *correct* object outcomes, is somewhat robust to small-to-moderate amounts of uncertainty during object subdialogues. The two leftmost bins in the figure, where COREF faces a mean uncertainty of between 1 and 2.3 states during an object subdialogue, represent the vast majority (95.3%) of the object subdialogues. This essentially corresponds to the tracking of one or two threads of interpretation during a subdialogue. A small percentage of the time, however, COREF faces a higher mean uncertainty during an object subdialogue, between 2.3 and 3 states, and this higher mean uncertainty seems to have a large negative impact on object outcomes.

In total, 13.1% of COREF’s *correct* object outcomes occur at a moment when COREF is uncertain what the state is (9.7% occur when COREF views two states as viable, and 3.4% occur when COREF views three contexts as viable). Thus, although maintaining a certain representation of a single thread of interpretation throughout a subdialogue is not strictly necessary for task success in COREF’s object identification game, uncertainty is clearly associated with problematic interactions.

5.1 Quantifying performance in collaborative ambiguity management

We now consider how effective COREF is at using ambiguity management questions (AMQs) to resolve its uncertainties. For the purposes of this analysis, we define an AMQ as a question asked by COREF at a moment when COREF was entertaining more than

² The distribution of the agent’s mean uncertainty is not normal. We therefore present our results in terms of the Wilcoxon rank sum test. For perspective, each Wilcoxon rank sum test reported in this section has also been compared to the results of a t-test, and the significance level always fell in the same general range.

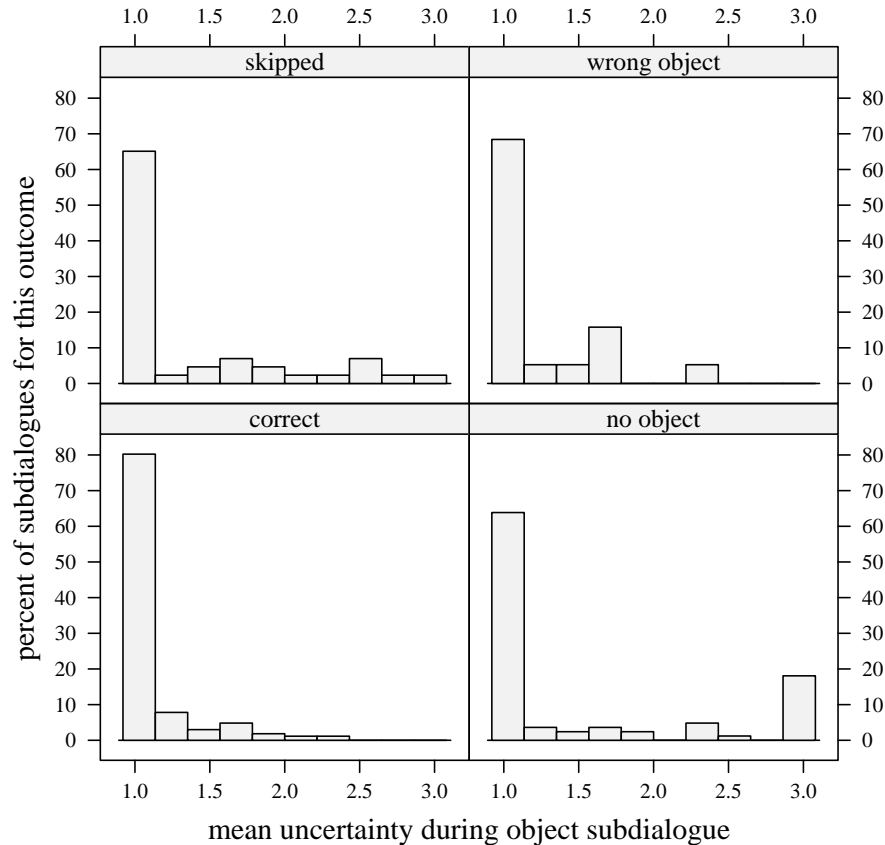


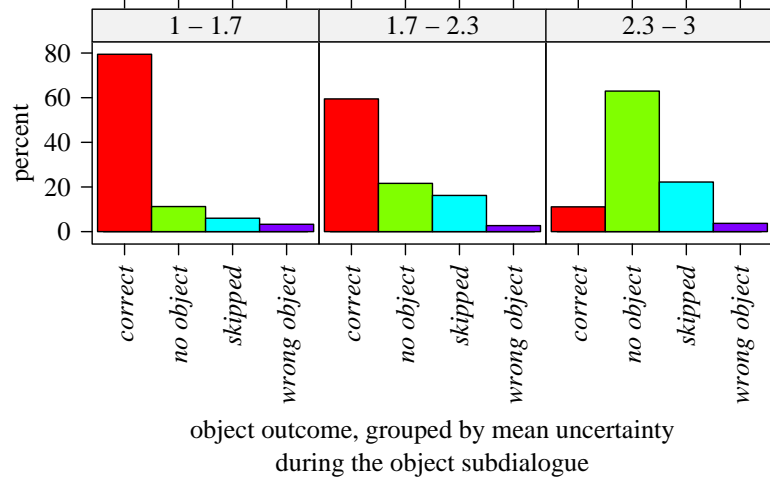
Figure 5

Mean uncertainty vs. object outcome during the object subdialogue. There are 435 objects (75.0%) in the *correct* bin, 83 objects (14.3%) in the *no object* bin, 43 objects (7.4%) in the *skipped* bin, and 19 objects (3.3%) in the *wrong object* bin.

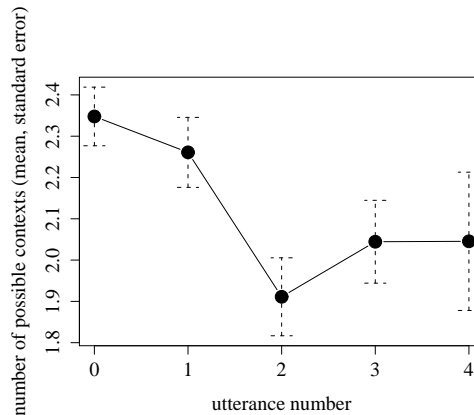
one thread of interpretation for the dialogue (or, equivalently, at a moment when COREF viewed more than one dialogue state as viable). Under this definition, several kinds of questions qualify as AMQs. These include explicit clarification questions, followups to ambiguous acknowledgments, and indeed any other yes/no question—even *is the target the beige circle?*—if COREF asks it while uncertain about the dialogue state. Our definition of AMQs thus reflects a general orientation towards question-asking under uncertainty, rather than towards specific categories of questions, such as clarifications.

In total, COREF asked 46 AMQs in the user study. This amounts to 2.3 AMQs per subject, on average. These 46 AMQs occurred during 41 (7.1%) of the 580 object subdialogues in the study. Thus, most object subdialogues did not include an AMQ. On a few occasions, COREF asked more than one AMQ during a single object subdialogue.

Figure 7 illustrates the effectiveness of COREF's question-asking policy at reducing uncertainty. As the figure shows, when COREF asks questions in an uncertain state, the mean reduction in the agent's uncertainty is about 0.4 states. The variation in uncertainty reduction reflects the fact that COREF's contribution tracking admits a number

**Figure 6**

Object outcome vs. mean uncertainty during the object subdialogue. There are 516 objects (88.97%) in the leftmost bin, 37 objects (6.38%) in the middle bin, and 27 objects (4.66%) in the rightmost bin.

**Figure 7**

Effect of ambiguity management questions on COREF's uncertainty. At utterance 0, COREF faces an ambiguity. At utterance 1, COREF has asked a question. Typically, at utterance 2, the user has answered COREF's question.

of different outcomes after a question is asked: the user's answer may eliminate one or more states; the user's answer may not eliminate any states; the user's answer may itself be ambiguous and *increase* the agent's uncertainty; the user may choose not to answer the question; or the agent may fail to understand the user's answer. COREF's contribution tracking allows it to execute a uniform update to its uncertainty for all of these different types of outcomes.

In Figure 7, the effect of COREF's AMQs can be approximately identified with the change in uncertainty that occurs between time 0, before COREF asks the question, and

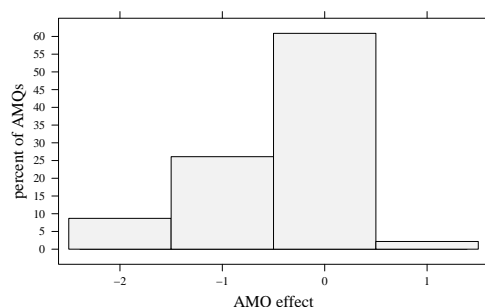


Figure 8
Distribution of AMQ effects for COREF's AMQs.

time 2, which is typically the time at which the user has just answered the question. We shall call this change in the number of viable contexts at time 0 and at time 2 the *AMQ effect*. Since COREF entertains exactly 1, 2, or 3 states as viable at each point in time, the AMQ effect $e = |v_2| - |v_0|$ necessarily lies in the range $-2 \leq e \leq 2$. A negative AMQ effect represents a reduction in the agent's uncertainty.

Figure 8 shows the distribution in AMQ effects for COREF's AMQs in this user study. The figure shows that the most common outcome, occurring for 62.2% of COREF's AMQs, is for no change in uncertainty to occur from time 0 to time 2. This largely reflects the frequent occurrence of *did you add it?* AMQs, to which the user generally responds *yes*. (When COREF asks *did you add it?*, the user has generally already added the object; however, even if they have not already added the object, the user usually tacitly adds the object and responds *yes*, rather than responding *no* to COREF's question.) In such AMQs, the agent's uncertainty is not reduced, but on the other hand, the agent does become certain that the object has been added, which allows it to decide to continue on to the next object with lower risk of a *no object* outcome.

For 35.6% of COREF's AMQs, the AMQ effect is -1 or -2, representing an elimination of 1 or 2 states, respectively. This category includes successful clarification questions such as the one in Figure 3. This analysis therefore allows us to quantify COREF's ability to successfully reduce its uncertainty by 1 or more states, when it decides to attempt to do so by asking an AMQ, at 35.6%.

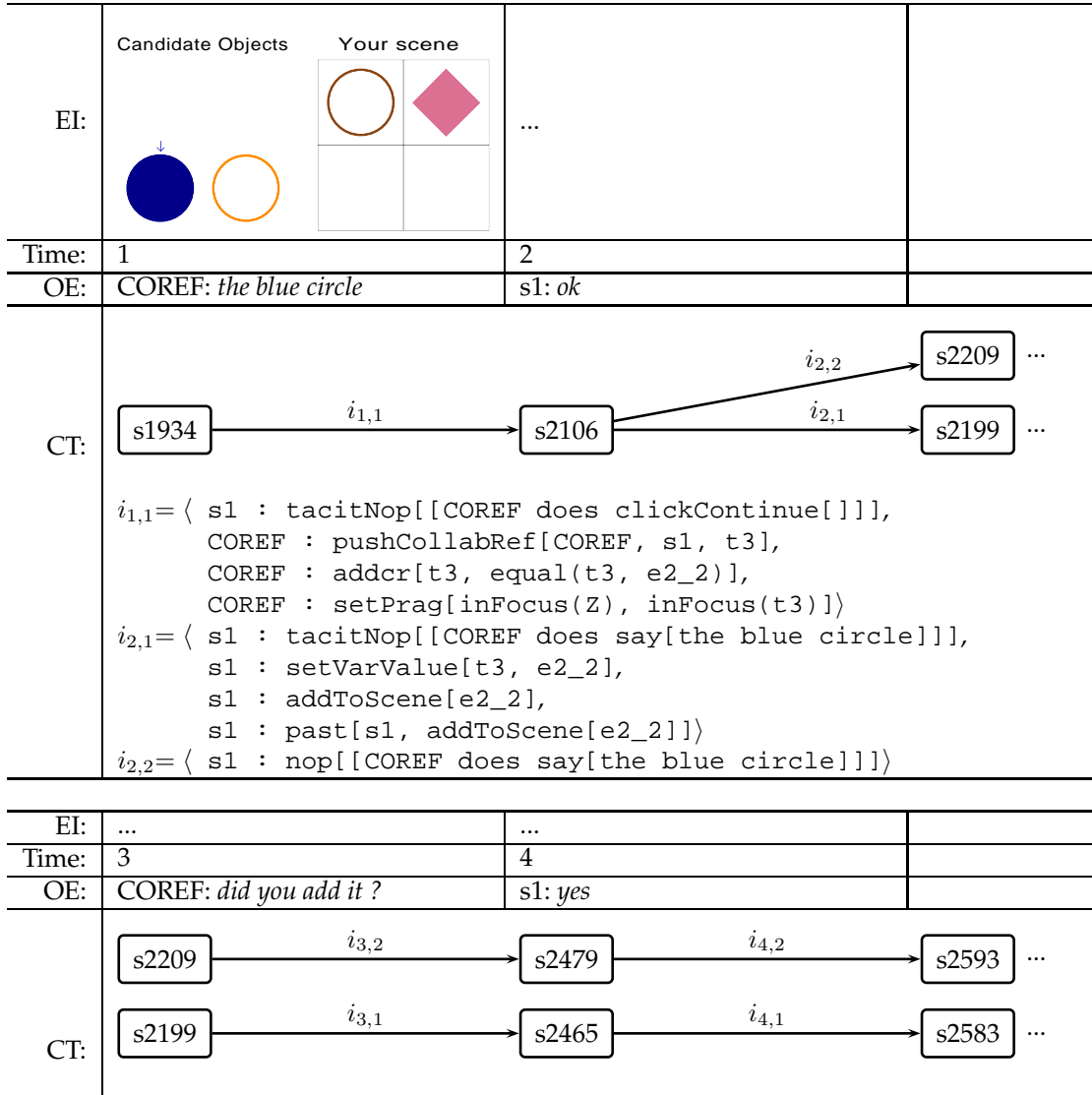
The remaining 2.2% of COREF's AMQs comprise a single AMQ whose effect was to increase COREF's uncertainty by 1 state.

Overall then, our analysis suggests that COREF's approach to ambiguity management is relatively successful in cases of mild or short-lived ambiguities. Such results can be compared to others presented for POMDP-based dialogue models. For example, Roy, Pineau, and Thrun (2000) investigate the reduction in the entropy in the system's belief state that occurs when the system asks a confirmation or clarification question following an ambiguous user utterance. What's new here is that our approach fully integrates COREF's ambiguity-related decision-making with its collaborative reasoning about language use in context. These results thus pave the way for applying the insights of probabilistic dialogue models within a range of frameworks for utterance interpretation and generation.

5.2 Capturing the ambiguity of acknowledgments

In Section 2, we motivated ambiguous acknowledgments as a challenge for incremental common ground models of context dependence and collaboration in language use. Contribution tracking allows COREF to capture these kinds of ambiguous acknowledgments using alternative threads of interpretation. Figure 9 provides one such example. COREF's reasoning in that example provides one possible model of the kind of reasoning that seems to be occurring in the human-human dialogue in Figure 2.

Figure 9: An attested subdialogue where COREF interprets an acknowledgment as ambiguous and proceeds.



continues on next page...

Figure 9: An attested subdialogue where COREF interprets an acknowledgment as ambiguous and proceeds. (continued)

$i_{3,1} = \langle$ COREF : tacitNop[[s1 does say[ok]]], COREF : pushRemind[COREF, s1, past, refuseTaskAction, addToScene[e2_2]], COREF : askYNQ[past[s1, addToScene[e2_2]]], COREF : setPrag[inFocus(Y), inFocus(e2_2)] $i_{3,2} = \langle$ COREF : tacitNop[[s1 does say[ok]]], s1 : setVarValue[t3, e2_2], COREF : pushRemind[COREF, s1, past, refuseTaskAction, addToScene[e2_2]], COREF : askYNQ[past[s1, addToScene[e2_2]]], COREF : setPrag[inFocus(Y), inFocus(e2_2)] $i_{4,1} = \langle$ s1 : tacitNop[[COREF does say[did you add it ?]]], s1 : past[s1, addToScene[e2_2]] $i_{4,2} = \langle$ s1 : tacitNop[[COREF does say[did you add it ?]]], s1 : addToScene[e2_2], s1 : past[s1, addToScene[e2_2]] \rangle			
EI:	
Time:	5	6	
OE:	[COREF does clickContinue[]]	[new reference problem]	
CT:	<pre> graph LR s2593 -- i5,2 --> s2712 s2583 -- i5,1 --> s2704 s2704 --- X[] s2712 -- i6,1 --> s2717 s2717 --- dots[...] </pre>		
	$i_{5,1} = \langle$ COREF : tacitNop[[s1 does say[yes]]], COREF : continueTask[t3] $i_{5,2} = \langle$ COREF : tacitNop[[s1 does say[yes]]], COREF : continueTask[t3] $i_{6,1} = \langle$ COREF : perceive[PerceivedNewReferenceProblemEvent<t4>] \rangle		

One feature worth special attention in Figure 9 is that the user’s response *yes* at time 4 is treated as ambiguous. The interpretation $i_{4,1}$ that fits the case where the user has already added the object to the scene simply shows the user passing on the chance to follow up COREF’s question and asserting in the affirmative. The second interpretation, $i_{4,2}$, by contrast, has the user first adding the object to the scene, as COREF has prompted them to do, and only then asserting that they have done so. The possibility of this reminder means that COREF remains uncertain about what actually happened in the dialogue—as reflected in the difference between s_{2583} and s_{2593} , and as noted already in our analysis of ambiguity management questions in Section 5.1. However, COREF is not uncertain about whether the user has identified the next object and placed it in the scene, so COREF is able here to continue with the task.

Because COREF is designed to collaborate under uncertainty about what the previous contributions have been, and because acknowledgments like *okay* are modeled as just another kind of contribution in COREF’s task models, COREF is relatively flexible

about how the ambiguity introduced by an ambiguous acknowledgment is resolved. For example, a user could in principle respond to *did you add it?* by saying *no* instead:

Example 1

COREF: the blue circle
 User: ok
 COREF: did you add it?
 User: no
 COREF: add it

In this dialogue, which shows COREF's implemented contribution tracking during a hypothetical variation on the observed dialogue in Figure 9, the user's response of *no* allows COREF to defuse its uncertainty and definitively interpret the user's previous utterance of *ok* as a simple acknowledgment. With its uncertainty resolved, COREF will simply proceed to issue a reminder to add the object.

An example from this user study in which COREF faces a richer kind of ambiguity about acknowledgments is this one:

Example 2

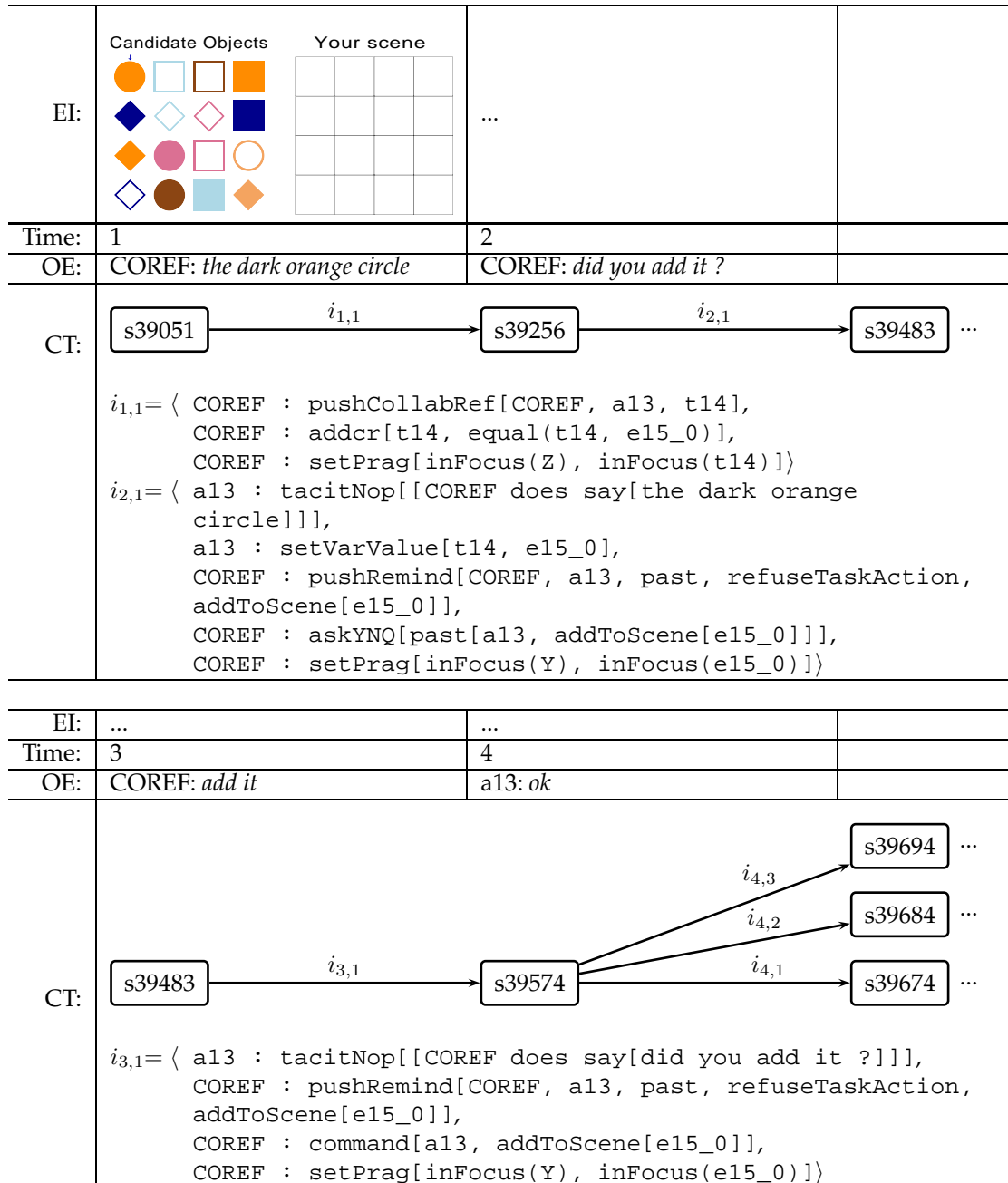
COREF: the dark orange circle
 COREF: did you add it?
 COREF: add it
 User: ok
 COREF: ok
 COREF: clicks continue to move to the next object

Figure 10 shows the dynamics of the COREF's state during this dialogue. When the user utters *ok* at time 4, COREF perceives the *ok* as ambiguous between three interpretations. According to $i_{4,3}$, the user simply acknowledges COREF's reminder to *add it*. According to $i_{4,1}$ and $i_{4,2}$, on the other hand, the user contributes that they have added the object to their scene.³ COREF therefore spawns three threads of interpretation. At time 5, COREF responds by saying *ok* back to the user, viewing itself as acknowledging the user's *ok* (no matter which contribution the user's *ok* actually made). Next, at time 6, COREF decides to click the `Continue (next object)` button and move on to the next object. In taking this action, COREF does something which is not coherent, according to its task models, under the assumption that the user's original *ok* was a simple acknowledgment (`nop`). (In fact, the user here had already added an object to their scene.) This lack of coherence means that COREF must drop that thread of interpretation at time 6. (A node depicted with a crosshatch pattern, such as `s40216` at time 6, indicates a state that COREF has made a strategic decision to "drop", or stop tracking.) This dynamic illustrates a new kind of

³ The kind of ambiguity that COREF perceives between $i_{4,1}$ and $i_{4,2}$, which differ in that $i_{4,2}$ substitutes a `tacitAbandonTasks` action for a `tacitNop` action, occurred frequently in this user study. Intuitively, we can explain this as COREF being unable to tell whether the user has abandoned the `Remind` task that COREF has previously pushed, or as instead tacitly acknowledging it. However, it is often hard to see exactly what the difference between these two particular interpretations is, and as such, we view this ambiguity as exhibiting a limitation in COREF's current interpretation process.

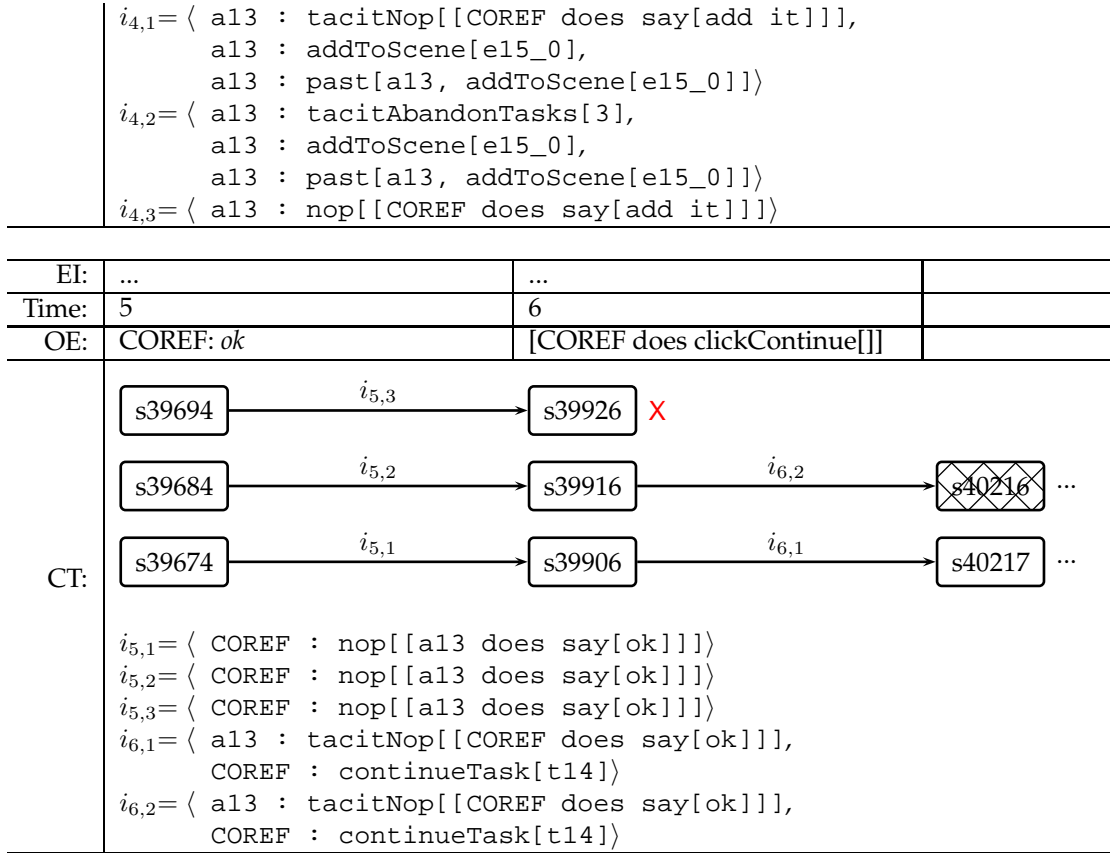
risk, as well as a new kind of flexibility, that contribution tracking affords to dialogue agents: they can assume a thread of coherent interpretation for the preceding dialogue that allows them to move forward with their task.

Figure 10: COREF proceeds despite an ambiguous acknowledgment



continues on next page...

Figure 10: COREF proceeds despite an ambiguous acknowledgment (continued)



In such examples, COREF therefore demonstrates a new qualitative capability for implemented dialogue systems that aim for collaborative language use: they can represent, reason about, and respond in collaborative ways to the ambiguity that accompanies acknowledgments in real-world, task-oriented dialogue.

5.3 Implementing ambiguity management as a flexible collaborative activity

Throughout this article, we have emphasized the problem solving that is at the heart of COREF's contribution tracking. Such problem-solving models bring a number of advantages.

For one thing, because COREF approaches the formulation of ambiguity management questions through problem solving, COREF synthesizes a wide variety of different utterances to achieve its goals in context. Figure 11 shows all the questions COREF asked in this user study.

Figure 11: All questions asked by COREF in the user study

count	question	count	question
159	<i>did you add it?</i>	21	<i>is the target round?</i>
12	<i>did you add the circle?</i>	18	<i>is the target solid?</i>
7	<i>did you add the square?</i>	16	<i>is the target square?</i>
6	<i>do you mean dark brown?</i>	2	<i>is the target the beige circle?</i>
1	<i>do you mean dark orange?</i>	2	<i>is the target the beige diamond?</i>
2	<i>do you mean it?</i>	2	<i>is the target the beige square?</i>
1	<i>do you mean the beige square?</i>	9	<i>is the target the blue circle?</i>
1	<i>do you mean the light blue circle?</i>	13	<i>is the target the blue diamond?</i>
1	<i>do you mean the light blue object?</i>	10	<i>is the target the blue square?</i>
1	<i>do you mean the pink circle?</i>	16	<i>is the target the brown circle?</i>
1	<i>do you mean the solid beige diamond?</i>	5	<i>is the target the brown diamond?</i>
1	<i>do you mean the solid circle?</i>	3	<i>is the target the brown square?</i>
1	<i>do you mean the solid pink circle?</i>	9	<i>is the target the circle?</i>
1	<i>do you mean the solid square?</i>	1	<i>is the target the dark blue circle?</i>
3	<i>i mean?</i>	2	<i>is the target the dark blue diamond?</i>
8	<i>is it beige?</i>	5	<i>is the target the dark blue square?</i>
1	<i>is it brown?</i>	1	<i>is the target the dark brown circle?</i>
5	<i>is it empty?</i>	1	<i>is the target the dark orange circle?</i>
2	<i>is it pink?</i>	2	<i>is the target the dark orange diamond?</i>
6	<i>is it round?</i>	9	<i>is the target the diamond?</i>
9	<i>is it solid?</i>	3	<i>is the target the empty brown circle?</i>
7	<i>is it square?</i>	2	<i>is the target the empty brown object?</i>
1	<i>is the circle an object?</i>	3	<i>is the target the empty circle?</i>
2	<i>is the circle dark blue?</i>	5	<i>is the target the empty diamond?</i>
1	<i>is the circle dark orange?</i>	1	<i>is the target the empty pink square?</i>
1	<i>is the circle light blue?</i>	4	<i>is the target the empty square?</i>
1	<i>is the circle the brown square?</i>	1	<i>is the target the light blue circle?</i>
1	<i>is the circle the dark blue square?</i>	1	<i>is the target the light blue diamond?</i>
1	<i>is the circle the light blue square?</i>	2	<i>is the target the light blue square?</i>
1	<i>is the circle the solid diamond?</i>	1	<i>is the target the orange circle?</i>
1	<i>is the diamond pink?</i>	1	<i>is the target the orange square?</i>
1	<i>is the diamond the object?</i>	1	<i>is the target the pink circle?</i>
1	<i>is the object the dark orange circle?</i>	5	<i>is the target the pink diamond?</i>
1	<i>is the square dark blue?</i>	1	<i>is the target the pink square?</i>
26	<i>is the target a diamond?</i>	2	<i>is the target the solid beige diamond?</i>
5	<i>is the target beige?</i>	1	<i>is the target the solid beige square?</i>
24	<i>is the target blue?</i>	1	<i>is the target the solid blue square?</i>
3	<i>is the target brown?</i>	1	<i>is the target the solid brown diamond?</i>
7	<i>is the target dark blue?</i>	4	<i>is the target the solid circle?</i>
8	<i>is the target dark brown?</i>	1	<i>is the target the solid dark blue circle?</i>
17	<i>is the target dark orange?</i>	1	<i>is the target the solid object?</i>
1	<i>is the target empty?</i>	1	<i>is the target the solid orange square?</i>
3	<i>is the target light blue?</i>	2	<i>is the target the solid square?</i>
2	<i>is the target pink?</i>	22	<i>is the target the square?</i>

COREF viewed all these utterances as contributive given the uncertainty (or certainty) it faced at the time it asked the question. The subtle variation in lexical choices between these questions reflects COREF's varying assessments of the expected recognizability of the contribution it wants to make. For example, a choice of *did you add it?* rather than *did you add the circle?* (or *did you add the square?*) generally depends on COREF's assessment about whether the object COREF wants to refer to is in focus in (all) the relevant context(s). A similar determination guides COREF's selection of *is it round?* rather than *is the target round?* in particular dialogue scenarios. The way COREF deploys such a wide variety of utterances is a natural outgrowth of its declarative representations and problem-solving mechanisms.

Another advantage of COREF's problem-solving approach to collaborative ambiguity management is that it provides a relatively flexible and robust approach to responding to perceived ambiguities. By **flexible**, we mean that disambiguating information can come in a variety of forms. By **robust**, we mean the information carried by an utterance is not forgotten, "downdated", or left in a pending status if an attempt to clarify fails. We now give a few examples that illustrate these features.

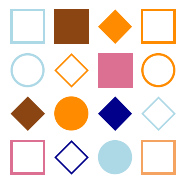
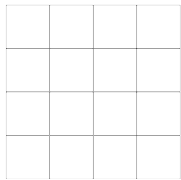
One outcome that one of COREF's AMQs can have is for the ambiguity to be completely resolved by the user's answer. We have seen several examples of this kind already. One example is the clarification question *do you mean dark brown?* which COREF asks in the interaction of Figure 3. Together with successful clarification questions, we can also include Example 1 as another type of interaction in which COREF is able to eliminate a perceived ambiguity using an AMQ; this time COREF is asking *did you add it?* rather than an explicit clarification question. These kinds of examples serve to make sense of COREF's representations, but they do not really showcase the advantages of modeling ambiguity resolution as extended collaboration.

However, an important qualitative feature of collaborative ambiguity management is that it allows an agent to keep talking and try to complete the task even when AMQs do not completely eliminate the ambiguity. For example, Figure 12 shows COREF's implemented contribution tracking during a hypothetical clarification attempt. In this interaction, the user begins by saying *the target is orange*. COREF sees the word *orange* as ambiguous, and responds with an AMQ: *do you mean dark orange?* The user here responds not by answering COREF's question, but rather by saying *it's an empty circle*. Given the array of visible objects, this response is actually a reasonably efficient way to complete the task: there is only one object that could be described as both *orange* and *an empty circle*. In this example, COREF is able to interpret the user's response *it's an empty circle* as abandoning COREF's clarification question, and adds two additional constraints – that the target is empty and a circle – as extensions to each of its two threads of interpretation. At time 4, by considering these threads, COREF is able to determine that only one thread of interpretation is internally consistent, and so COREF adds the dark orange empty circle to its scene and instructs the user to continue on to the next object. COREF's ambiguity is resolved. An important aspect of this example is that COREF needs to aggregate information both from the ambiguous utterance and from information that follows its failed clarification attempt in order to complete the task successfully.

Figure 12: COREF's uncertainty is resolved flexibly

continues on next page...

Figure 12: COREF's uncertainty is resolved flexibly (continued)

EI:	<div style="display: flex; justify-content: space-around;"> <div style="text-align: center;"> <p>Candidate Objects</p>  </div> <div style="text-align: center;"> <p>Your scene</p>  </div> </div>	...	
Time:	1	2	
OE:	User: <i>the target is orange</i>	COREF: <i>do you mean dark orange ?</i>	
CT:	<div style="display: flex; justify-content: space-between; align-items: center;"> <div style="border: 1px solid black; padding: 5px;">s8822</div> <div style="text-align: center;"> $i_{1,2}$ $i_{1,1}$ </div> <div style="border: 1px solid black; padding: 5px;">s12025</div> <div style="text-align: center;"> $i_{2,2}$ $i_{2,1}$ </div> <div style="border: 1px solid black; padding: 5px;">s12240</div> <div style="text-align: center;">...</div> </div> <div style="display: flex; justify-content: space-between; align-items: center; margin-top: 10px;"> <div style="border: 1px solid black; padding: 5px;">s12010</div> <div style="border: 1px solid black; padding: 5px;">s12230</div> <div style="text-align: center;">...</div> </div> <p style="margin-top: 20px;"> $i_{1,1} = \langle$ User : pushCollabRef[User, COREF, t14], User : addcr[t14, darkorangeFigureObject(t14)], User : setPrag[inFocus(Y), inFocus(t14)] \rangle </p> <p> $i_{1,2} = \langle$ User : pushCollabRef[User, COREF, t14], User : addcr[t14, sandybrownFigureObject(t14)], User : setPrag[inFocus(Y), inFocus(t14)] \rangle </p> <p> $i_{2,1} = \langle$ COREF : pushClarify[[User does say[the target is orange]], User, COREF, Orange], COREF : pushYNQ[COREF, User, addcr[Orange, P], negcr[Orange, P]], COREF : askYNQ[addcr[Orange, equal(Orange, darkorangeFigureObject)]] \rangle </p> <p> $i_{2,2} = \langle$ COREF : pushClarify[[User does say[the target is orange]], User, COREF, Orange], COREF : pushYNQ[COREF, User, addcr[Orange, P], negcr[Orange, P]], COREF : askYNQ[addcr[Orange, equal(Orange, darkorangeFigureObject)]] \rangle </p>		
EI:	
Time:	3	4	
OE:	User: <i>it's an empty circle</i>	COREF: <i>click continue</i>	
CT:	<div style="display: flex; justify-content: space-between; align-items: center;"> <div style="border: 1px solid black; padding: 5px;">s12240</div> <div style="text-align: center;"> $i_{3,2}$ </div> <div style="border: 1px solid black; padding: 5px;">s18528</div> <div style="text-align: center;">X</div> </div> <div style="display: flex; justify-content: space-between; align-items: center; margin-top: 10px;"> <div style="border: 1px solid black; padding: 5px;">s12230</div> <div style="text-align: center;"> $i_{3,1}$ </div> <div style="border: 1px solid black; padding: 5px;">s18509</div> <div style="text-align: center;"> $i_{4,1}$ </div> <div style="border: 1px solid black; padding: 5px;">s18679</div> <div style="text-align: center;">...</div> </div>		

continues on next page...

Figure 12: COREF's uncertainty is resolved flexibly (continued)

```

i3,1 = { User : tacitAbandonTasks[4],
         User : addcr[t14, circleFigureObject(t14)],
         User : setPrag[inFocus(Y), inFocus(t14)],
         User : addcr[t14, emptyFigureObject(t14)] }
i3,2 = { User : tacitAbandonTasks[4],
         User : addcr[t14, circleFigureObject(t14)],
         User : setPrag[inFocus(Y), inFocus(t14)],
         User : addcr[t14, emptyFigureObject(t14)] }
i4,1 = { COREF : tacitNop[[User does say[it's an empty circle]]],
         COREF : setVarValue[t14, e14_0],
         COREF : addToScene[e14_0],
         COREF : pushRemind[COREF, User, past, refuseTaskAction,
                             continueTask[t14]],
         COREF : command[User, continueTask[t14]] }

```

Another reason it can be advantageous to defer the resolution of a perceived ambiguity is that the user may disambiguate their own intention without specific intervention by the system. For example, Figure 13 shows COREF's implemented contribution tracking during a hypothetical interaction that illustrates this possibility. In this interaction, the user starts out by saying *a blue object*, which COREF perceives as ambiguous due to the presence of objects of two shades of blue in the display. Before COREF responds, the user continues by saying *the dark blue one*. This effectively eliminates COREF's perceived ambiguity without any specific effort from COREF.

Figure 13: COREF's uncertainty is resolved flexibly

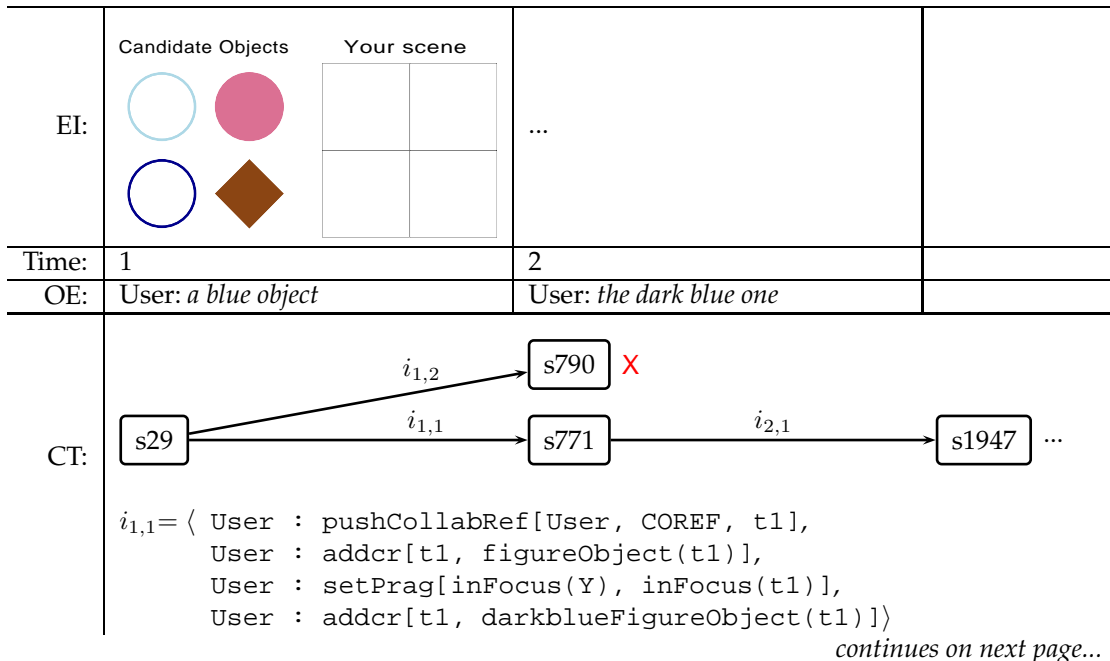


Figure 13: COREF's uncertainty is resolved flexibly (continued)

	$i_{1,2} = \langle$ User : pushCollabRef[User, COREF, t1], User : addcr[t1, figureObject(t1)], User : setPrag[inFocus(Y), inFocus(t1)], User : addcr[t1, lightblueFigureObject(t1)] $i_{2,1} = \langle$ COREF : tacitNop[[User does say[a blue object]]], User : addcr[t1, equal(t1, e3_0)], User : setPrag[inFocus(Z), inFocus(t1)]
EI:	...
Time:	3
OE:	COREF: <i>click continue</i>
CT:	<div style="display: flex; align-items: center; gap: 20px;"> <div style="border: 1px solid black; padding: 2px 5px;">s1947</div> <div style="text-align: center;"> $\xrightarrow{i_{3,1}}$ </div> <div style="border: 1px solid black; padding: 2px 5px;">s2070</div> <div>...</div> </div> $i_{3,1} = \langle$ COREF : tacitNop[[User does say[the dark blue one]]], COREF : setVarValue[t1, e3_0], COREF : addToScene[e3_0], COREF : pushRemind[COREF, User, past, refuseTaskAction, continueTask[t1]], COREF : command[User, continueTask[t1]]

Taken together, these examples demonstrate a new qualitative capability for dialogue systems: we can use contribution tracking to implement ambiguity management as a flexible collaborative activity. This approach may help us to design agents that can keep talking and resolve their uncertainty over time, as needed.

6. Conclusion

We have presented a framework that allows task-oriented dialogue agents to use language collaboratively despite uncertainty about the dialogue state. We have presented empirical evidence that managing ambiguity is a key task for dialogue agents such as ours, and that it can be addressed successfully as a general problem of collaboration under uncertainty. In particular, our model shows how dialogue agents can support grounding acknowledgments, clarification of ambiguous utterances, and task-oriented question asking using generic linguistic resources and goal-oriented ambiguity management strategies.

Our model is provisional in many ways. We have said relatively little, for example, about the problem solving involved in deciding whether a potential contribution would make progress in the task. In fact, COREF's simple policy sometimes proposes clarification questions whose answers would not always resolve COREF's ambiguities. An example is *do you mean it?*—COREF uses the utterance because only one object at once is the salient referent for *it*, so the utterance makes a unique acceptable move in each of the possible contexts. In a sense, COREF is right: this question is not ambiguous in a certain context; neither is its expected answer, *yes*. We overlooked the constraint that clarifying moves should elicit different behaviors in different alternative states. We

take such cases as evidence of a general need in contribution tracking for better models of collaboration under uncertainty.

Our representations themselves are also quite provisional. We believe our model could be extended from clarification of ambiguities in semantics and pragmatics, as we have explored here, to clarifications of perceived ambiguities at other levels of linguistic representation, drawing on the work of Ginzburg and Cooper (2004). In principle, perceived phonologic or syntactic ambiguities could be translated into ambiguities in the context resulting from an utterance, entirely analogously to COREF's response to ambiguities of meaning. It would not necessarily be easy to demonstrate this, however. Working with the output of a generic speech recognizer or parser, for example, would involve a radical proliferation of ambiguity and would require much more sophisticated inference strategies and implementation techniques.

Meanwhile, our work does not immediately cover clarification questions that are not designed to resolve perceived ambiguities, but rather are asked in situations where *no* interpretations are found. Such examples occur; see Ginzburg and Cooper (2004) or Purver (2004) for examples. When COREF finds no interpretations for a user utterance, it notes the utterance and signals an interpretation failure (currently by saying *umm*), but it otherwise leaves its context representation as it was, and is unable to address the failure with its usual ambiguity management policy. Alternative characterizations of agents' reasoning in such cases are still required, and work such as Purver's provides a natural starting point.

Traditional classifications of grounding actions (Traum 1999) include a variety of other cases that we do not handle. For example, we do not treat repair requests like *what?* or *what did you say?*, which can signal interpretation failure or the hearer's incredulity at the speaker's apparent (but correctly and uniquely identified) meaning. Similarly, we do not treat self-repairs by speakers. These can exclude a possible but unintended interpretation, to avoid a foreseen misunderstanding—an example in COREF's domain would be, *A: I moved it. A: I mean I moved the blue circle*. They can also correct a prior verbal mistake, as when a speaker has mistakenly used the wrong word: *A: I moved the circle. A: I mean I moved the square*. It would be interesting to explore whether richer models of domain uncertainty and dialogue context would enable us to account for these utterance types.

Thus, much further work is required to scale these techniques to richer domains and more capable dialogue systems. Nevertheless, we believe that these results showcase how judicious system-building efforts can lead to dialogue capabilities that defuse some of the bottlenecks to robust natural language interaction. In particular, a focus on improving our agents' basic abilities to tolerate and resolve ambiguities as a dialogue proceeds may prove to be a valuable technique for improving the overall dialogue competence of the agents we build.

7. Acknowledgments

This work was sponsored in part by NSF CCF-0541185 and HSD-0624191, and by the U.S. Army Research, Development, and Engineering Command (RDECOM). Statements and opinions expressed do not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred. Thanks to our reviewers, Rich Thomason, David Traum and Jason Williams.

References

- Allen, James, Nathanael Chambers, George Ferguson, Lucian Galescu, Hyuckchul Jung, Mary Swift, and William Tayson. 2007. PLOW: A collaborative task learning agent. In *National Conference on Artificial Intelligence (AAAI)*.
- Asher, N. and A. Lascarides. 2003. *Logics of Conversation*. Cambridge.
- Blaylock, Nate, James Allen, and George Ferguson. 2002. Managing communicative intentions with collaborative problem solving. In Ronnie Smith and Jan van Kuppevelt, editors, *Current and New Directions in Dialogue*. Kluwer.
- Bohus, D. and A. Rudnicky. 2006. A k hypotheses + other belief updating model. In *Proceedings AAAI Workshop on Statistical and Empirical Approaches for Spoken Dialogue Systems*.
- Brennan, Susan E. and Herbert H. Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology*, 22(6):1482–1493.
- Bunt, H. 2000. Dialogue pragmatics and context specification. In H. Bunt and W. Black, editors, *Abduction, Belief and Context in Dialogue*. Benjamin, pages 81–150.
- Camerer, Colin F. 2003. Behavioral studies of strategic thinking in games. *Trends in Cognitive Sciences*, 7(5):225–231.
- Carberry, Sandra. 2001. Techniques for plan recognition. *User Modeling and User-Adapted Interaction*, 11:31–48.
- Clark, H. H. and E. F. Schaefer. 1989. Contributing to discourse. *Cognitive Science*, 13:259–294.
- Clark, Herbert. 1996. *Using Language*. Cambridge University Press, Cambridge, UK.
- Clark, Herbert H. and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. In Philip R. Cohen, Jerry Morgan, and Martha E. Pollack, editors, *Intentions in Communication*. MIT Press, Cambridge, Massachusetts, 1990, pages 463–493.
- Cohen, Philip R. and Hector J. Levesque. 1990. Intention is choice with commitment. *Artificial Intelligence*, 42:213–261.
- Core, Mark G. and James F. Allen. 1997. Coding dialogues with the DAMSL annotation scheme. In David Traum, editor, *Working Notes: AAAI Fall Symposium on Communicative Action in Humans and Machines*, pages 28–35, Menlo Park, California. American Association for Artificial Intelligence.
- DeVault, David. 2008. *Contribution Tracking: Participating in Task-Oriented Dialogue under Uncertainty*. Ph.D. thesis, Department of Computer Science, Rutgers, The State University of New Jersey, New Brunswick, NJ.
- DeVault, David, Natalia Kariaeva, Anubha Kothari, Iris Oved, and Matthew Stone. 2005. An information-state approach to collaborative reference. In *ACL 2005 Proceedings Companion Volume. Interactive Poster and Demonstration Sessions*, pages 1–4, University of Michigan.
- DeVault, David, Charles Rich, and Candace L. Sidner. 2004. Natural language generation and discourse context: Computing distractor sets from the focus stack. In *Proceedings of the 17th International Florida Artificial Intelligence Research Society Conference (FLAIRS 2004)*, pages 887–892.
- DeVault, David and Matthew Stone. 2006. Scorekeeping in an uncertain language game. In *Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue (SemDial-10)*, pages 139–146.
- DeVault, David and Matthew Stone. 2007. Managing ambiguities across utterances in dialogue. In *Proceedings of the 11th Workshop on the Semantics and Pragmatics of Dialogue (Decalog 2007)*, pages 49–56.
- DeVault, David and Matthew Stone. 2009. Learning to interpret utterances using dialogue history. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL)*, pages 184–192.
- Ginzberg, Jonathan and Robin Cooper. 2004. Clarification, ellipsis and the nature of contextual updates in dialogue. *Linguistics and Philosophy*, 27(3):297–365.
- Ginzburg, Jonathan and Robin Cooper. 2004. Clarification, ellipsis and the nature of contextual updates in dialogue. *Linguistics and Philosophy*, 27.
- Grosz, Barbara J. and Sarit Kraus. 1996. Collaborative plans for complex group action. *Artificial Intelligence*, 86(2):269–357.
- Grosz, Barbara J. and Candace L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.
- Heeman, Peter. 2007. Combining reinforcement learning with information-state update rules. In *NAACL*, pages 268–275.

- Heeman, Peter A. and Graeme Hirst. 1995. Collaborating on referring expressions. *Computational Linguistics*, 21(3):351–383.
- Henderson, James, Oliver Lemon, and Kalliroi Georgila. 2008. Hybrid reinforcement/supervised learning of dialogue policies from fixed data sets. *Computational Linguistics*, 34(4):487–511.
- Hobbs, Jerry R., Mark Stickel, Douglas Appelt, and Paul Martin. 1993. Interpretation as abduction. *Artificial Intelligence*, 63:69–142.
- Horvitz, Eric and Tim Paek. 2001. Harnessing models of users' goals to mediate clarification dialog in spoken language systems. In *Proceedings of the Eighth International Conference on User Modeling*, pages 3–13.
- Jaeger, Gerhard. 2008. Applications of game theory in linguistics. *Language and Linguistics Compass*, 2(3):406–421.
- Janarthanam, Srini and Oliver Lemon. 2009. Learning lexical alignment policies for generating referring expressions for spoken dialogue systems. In *EACL*.
- Larsson, S. and D. Traum. 2000. Information state and dialogue management in the TRINDI dialogue move engine toolkit. *Natural Language Engineering*, 6:323–340.
- Lemon, Oliver, Kalliroi Georgila, James Henderson, and Matthew Stuttle. 2006. An ISU dialogue system exhibiting reinforcement learning of dialogue policies: generic slot-filling in the TALK in-car system. In *EACL*.
- Lesh, Neal, Charles Rich, and Candace L. Sidner. 2001. Collaborating with focused and unfocused users under imperfect communication. In M. Bauer, P. J. Gmytrasiewicz, and J. Vassileva, editors, *User Modeling (UM)*, volume 2109 of *LNAI*, pages 64–73.
- Lewis, David. 1979. Score-keeping in a language game. In Paul Portner and Barbara H. Partee, editors, *Formal Semantics: The Essential Readings*. Blackwell, Oxford, UK, pages 162–177.
- Litman, Diane J. and James F. Allen. 1987. A plan recognition model for subdialogues in conversation. *Cognitive Science*, 11:163–200.
- Lochbaum, Karen E. 1998. A collaborative planning model of intentional structure. *Computational Linguistics*, 24(4):525–572.
- Matheson, C., M. Poesio, and D. Traum. 2000. Modelling grounding and discourse obligations using update rules. In *Proceedings of NAACL*, May.
- Paek, Tim and Roberto Pieraccini. 2008. Automating spoken dialogue management design using machine learning: An industry perspective. *Speech Communication*, 50(8–9):716–729.
- Perrault, C. Raymond, James F. Allen, and Philip R. Cohen. 1978. Speech acts as a basis for understanding dialogue coherence. In *Proceedings of the 1978 workshop on Theoretical issues in natural language processing*, pages 125–132, Morristown, NJ, USA. Association for Computational Linguistics.
- Poesio, Massimo and David R. Traum. 1997. Conversational actions and discourse situations. *Computational Intelligence*, 13(3):309–347.
- Pollack, Martha E. 1990. Plans as complex mental attitudes. In Philip Cohen, Jerry Morgan, and Martha Pollack, editors, *Intentions in Communication*. MIT Press, Cambridge, Massachusetts, pages 77–103.
- Pollack, Martha E. 1992. The uses of plans. *Artificial Intelligence*, 57:43–68.
- Power, Richard. 1977. The organisation of purposeful dialogues. *Linguistics*, 17:107–152.
- Purver, Matthew. 2004. *The Theory and Use of Clarification Requests in Dialogue*. Ph.D. dissertation, Department of Computer Science, King's College, University of London, London.
- Rich, Charles, Candace L. Sidner, and Neal Lesh. 2001. Collagen: Applying collaborative discourse theory to human-computer interaction. *Artificial Intelligence Magazine*, 22(4):15–25.
- Rieser, Verena and Oliver Lemon. 2008. Learning effective multimodal dialogue strategies from Wizard-of-Oz data: Bootstrapping and evaluation. In *Proceedings of ACL-08: HLT*, pages 638–646, Columbus, Ohio, June. Association for Computational Linguistics.
- Rieser, Verena and Oliver Lemon. 2009. Natural language generation as planning under uncertainty for spoken dialogue systems. In *EACL*.
- Roy, Nicholas, Joelle Pineau, and Sebastian Thrun. 2000. Spoken dialog management for robots. In *The Proceedings of the Association for Computational Linguistics*.
- Simon, Herbert A. 1978. Rationality as process and as product of thought. *The American Economic Review*, 68(2):1–16.
- Stalnaker, Robert. 1974. Pragmatic presuppositions. In Robert Stalnaker, editor, *Context and Content*. Oxford, New York, New York, pages 47–62.
- Stone, Matthew. 2002. Lexicalized grammar 101. In *ACL Workshop on Tools and Methodologies for Teaching Natural Language Processing*.

- Stone, Matthew. 2004a. Communicative intentions and conversational processes in human-human and human-computer dialogue. In Trueswell and Tanenhaus, editors, *World-Situated Language Use*. MIT.
- Stone, Matthew. 2004b. Intention, interpretation and the computational structure of language. *Cognitive Science*, 28(5):781–809.
- Stone, Matthew, Christine Doran, Bonnie Webber, Tonia Bleam, and Martha Palmer. 2003. Microplanning with communicative intentions: the spud system. *Computational Intelligence*, 19(4):314–381.
- Thomason, Richmond. 1990. Accommodation, meaning, and implicature: Interdisciplinary foundations for pragmatics. In Philip R. Cohen, Jerry Morgan, and Martha Pollack, editors, *Intentions in Communication*. MIT Press, Cambridge, Massachusetts, pages 325–363.
- Thomason, Richmond H., Matthew Stone, and David DeVault. 2006. Enlightened update: A computational architecture for presupposition and other pragmatic phenomena. For the Ohio State Pragmatics Initiative, 2006, available at <http://www.research.rutgers.edu/~ddevault/>.
- Traum, David R. 1994. *A Computational Theory of Grounding in Natural Language Conversation*. Ph.D. dissertation, Department of Computer Science, University of Rochester, Rochester, New York.
- Traum, David R. 1999. Computational models of grounding in collaborative systems. In Susan E. Brennan, Alain Giboin, and David Traum, editors, *Working Papers of the AAAI Fall Symposium on Psychological Models of Communication in Collaborative Systems*, pages 124–131, Menlo Park, California. American Association for Artificial Intelligence, American Association for Artificial Intelligence.
- Webber, Bonnie, Matthew Stone, Aravind Joshi, and Alistair Knott. 2003. Anaphora and discourse structure. *Computational Linguistics*, 29(4):545–588.
- Williams, Jason and Steve Young. 2007. Partially observable markov decision processes for spoken dialog systems. *Computer Speech and Language*, 21(2):393–422.