1 Pursuing and Demonstrating Understanding in Dialogue

David DeVault and Matthew Stone
University of Southern California and Rutgers University

1.1 Introduction

The appeal of dialogue as an interface modality is its ability to support open-ended mixed-initiative interaction. Many systems offer rich and extensive capabilities, but must support infrequent and untrained users. In such cases, it's unreasonable to expect users to know the actions they need in advance, or to be able to specify them using a regimented scheme of commands or menu options. Dialogue offers the potential for the user to talk through their needs with the system and arrive collaboratively at a feasible solution.

Dialogue, in short, comes into its own in potentially problematic interactions. We do not expect the user's conceptualizations of the task and domain to align with the system's. The system cannot count on some fixed regime to recover the meanings of the user's words, their reference in the domain, or their contribution to ongoing activity. The system must be prepared for incorrect or incomplete analyses of users' utterances, and must be able to pinpoint users' needs across extended interactions. Conversely, the system must be prepared for users that misunderstand it, or fail to understand it. This chapter provides an overview of the concepts, models and research challenges involved in this process of pursuing and demonstrating understanding in dialogue.

We start in Section 1.2 from analyses of human–human conversation. People are no different from systems: they face potentially problematic interactions whenever they must sort out issues that some find unfamiliar. In response, they avail themselves of a wide range of discourse moves and interactive strategies, suggesting that they approach communication itself as a collaborative process of agreeing, to their mutual satisfaction, on the distinctions that matter for their discussion and on the expressions through which to identify those distinctions. In the literature, this process is often described as *grounding* communication, or identifying contributions well enough so that they become part of the *common ground* of the conversation [Clark and Marshall, 1981, Clark and Wilkes-Gibbs, 1986, Clark and Schaefer, 1989, Clark, 1996].

For a computer system, grounding can naturally be understood in terms of problem solving. When a system encounters an utterance whose interpretation is incomplete, ambiguous, unlikely or unreliable, it has to figure out how to refine and confirm that interpretation without derailing the interaction. When a sys-

tem gets evidence that one of its own utterances may not have been understood correctly, it has to figure out how to revise and reframe its contributions to keep the conversation on track. Solving such problems means coming up with ways building on partial understandings of previous contributions, while formulating utterances with a reasonable expectation that they will be understood correctly no matter what, and will move the conversation forward. We call this reasoning process *contribution tracking*. It is a core requirement for natural language generation in dialogue.

In Section 1.3, we describe a preliminary version of contribution tracking in a prototype dialogue system, COREF [DeVault and Stone, 2007, 2006, DeVault et al., 2005], and offer a new assessment of the qualitative capabilities that contribution tracking gives the current version of COREF. We emphasize how contribution tracking influences all the steps of planning and acting in dialogue systems, but particularly the process of natural language generation. Our analysis suggests that dialogue systems designers must be wary of idealizations often adopted in natural language generation research. For example, you may not be able to specify a definite context for generation, you may not be able to formulate an utterance you can guarantee that the user will understand, and you may have a communicative goal and content to convey that overlaps in important ways with the communicative goals and content of previously-formulated utterances.

Relaxing these idealizations remains a major challenge for generation and grounding in dialogue. In particular, in Section 1.4, we use our characterization of contribution tracking as problem solving to analyze the capabilities of grounding models in current conversational systems. Many applied frameworks restrict grounding actions to simple acknowledgments, confirmation utterances and clarification requests. This narrow focus lets systems streamline the reasoning required to generate grounding utterances, through approaches that abstract away from aspects of the system's uncertainty about the conversational state and its consequences for system choices.

On the other hand, as we survey in Section 1.5, a wide range of emerging applications will require a much more sophisticated understanding of the grounding work that systems and users are doing. A conversational agent that generates communicative gestures and facial expressions will need to model how nonverbal actions help to signal understanding or lack of understanding. A collaborative robot that carries out physical activities jointly with a human partner will need to model how real-world actions give evidence of interlocutors' interpretations of utterances. A virtual human, with naturalistic mechanisms of attention, cognition and emotion, will need to recognize that its internal state, including its understanding of what is happening in the conversation, is often legible in the work it is obviously doing to participate in the conversation. In our view, these emerging applications for dialogue technology give a lasting importance to general accounts of grounding as problem solving—and offer an exciting range of practical test cases for generating new kinds of grounding phenomena.

1.2 Background

Avoiding misunderstanding in problematic situations is a joint effort. If the addressee gives no feedback about what he understands, there is no way that the speaker can confirm the she was understood as she intended. Conversely, unless the speaker acts on the feedback the addressee provides, the addressee cannot correct an incomplete or faulty understanding. In human–human conversations, interlocutors do work jointly this way to stay in synch. Understanding of what people do in the face of difficulties can provide a starting point for achieving similar grounding in dialogue systems.

1.2.1 Grounding behaviors

As Clark [1996, Ch 5] reminds us, successful understanding, for humans and machines, involves recognizing what the speaker is doing across a hierarchy of levels. At the lowest level, recognizing words, people sometimes face substantial difficulty—though perhaps not so severe as systems with automatic speech recognition. In such situations, hearers often use confirmation strategies to make sure they get the information right. Example (1.1) illustrates:

(1.1) June: ah, what ((are you)) now, *where*
Darryl: *yes* forty-nine Skipton Place
June: forty-one
Darryl: nine. nine
June: forty-nine, Skipton Place,
Darryl: W one.
London–Lund Corpus (9.2a.979) cited in [Clark, 1996, p. 236]

June repeats what Darryl says. Note how this enables Darryl to catch and correct June's mishearing of *forty-nine* as *forty-one*. Notice also the coordination involved. Darryl speaks in installments that can be echoed and corrected easily; June echoes as expected. The strategy allows specifications, confirmations and corrections to come fluidly and elliptically. It's not just June that's taking grounding into account in managing the dialogue here; it's also Darryl.

Understanding at the next-higher level involves the grammatical analysis of the utterance. A representative task here is the resolution of word-sense ambiguities. Systems famously face a vocabulary problem because users are so variable in the meanings they assign to words in unfamiliar situations [Furnas et al., 1987]. So do people, like interlocutors A and B in (1.2):

(1.2) B: k– who evaluates the property —
A: u:h whoever you asked, . the surveyor for the building society
B: no, I meant who decides what price it'll go on the market —
A: (– snorts) , whatever people will pay —
London–Lund Corpus (4.2.298) cited in [Clark, 1996, p. 234]

In the complex process of valuing real estate, property is evaluated in one sense by the seller and their sales agent to fix an offering price, and evaluated in another sense by an appraisal or survey carried out before the buyer can get a mortgage, which in the UK prototypically comes from a building society. Interestingly in (1.2), A simply proceeds to answer B, assuming one construal of B's term *evaluate*. Even though A offers no overt confirmation or acknowledgment of B's meaning, the response allows B to recognize A's different construal and to reframe the original question more precisely. In this case, grounding is accomplished through explicit meta-level language that takes meaning itself as the topic of conversation in grounding episodes.

The highest level of understanding concerns the relationship of interlocutors' meanings to the ongoing task and situation. Problematic reference, as in (1.3), illustrates the difficulties both people and systems face, and the dynamics through which people achieve grounding.

(1.3) A: Okay, the next one is the rabbit.
B: Uh —
A: That's asleep, you know, it looks like it's got ears and a head pointing down?
B: Okay

[Clark and Wilkes-Gibbs, 1986] cited in [Clark, 1993, p. 127]

Here B offers a simple signal of non-understanding. At this point it is up to A to develop the initial description further. Here A produces an an *expansion*, in Clark and Wilkes-Gibbs's [1986] terminology. A ignores B's possible invitation to address the explicit topic of what A means, and simply provides a syntactic and semantic continuation of the initial description.

In our characterization of grounding so far, we've seen that an addressee's contribution to grounding can include confirmation, relevant followup utterances, and signals of non-understanding. We close with two other examples that underscore the range of grounding moves in human–human conversation. In example (1.4), the addressee responds to an unclear description by offering an alternative description that seems clearer.

(1.4) A: Okay, and the next one is the person that looks like they're carrying something and it's sticking out to the left. It looks like a hat that's upside down.
B: The guy that's pointing to the left again?
A: Yeah, pointing to the left, that's it! (laughs)
B: Okay

[Clark and Wilkes-Gibbs, 1986] cited in [Clark, 1993, p. 129]

The speaker accepts and adopts the addressee's reformulation. Such cases show the benefits of joint effort in grounding.

Finally, we shouldn't forget simple cases like (1.5):

(1.5) Burton: how was the wedding —
Anna: oh it was really good, it was uh it was a lovely day
Burton: yes
Anna: and . it was a super place, . to have it . of course
Burton: yes —
Anna: and we went and sat on sat in an orchard, at Grantchester, and had a huge tea *afterwards (laughs —)*
Burton *(laughs —)*
London–Lund corpus (7.3l.1441) cited in [Clark, 1996, p. 237]

Grounding is necessary even in unproblematic interactions, and it often takes the form of straightforward acknowledgments, like Burton's in (1.5), where addressees simply indicate their judgment that they understand what the speaker has contributed so far.

1.2.2 Grounding as a collaborative process

Clearly, grounding in human–human conversation is a complex, wide-ranging skill. Its effects are pervasive, multifaceted and varied. Implementing analogous skills in dialogue systems requires an overarching framework that reveals the fundamental commonalities behind people's grounding strategies and links them to mechanisms that plausibly underlie them.

Accounts of grounding in cognitive science start from Grice's influential account of conversation as rational collaboration [Grice, 1975]. Grice proposes that conversation is governed by a central Cooperative Principle: all interlocutors are expected do their part in the conversation, by making appropriate contributions. This includes showing how they understand previous utterances in the conversation and following up utterances to ensure understanding [Clark and Wilkes-Gibbs, 1986, Clark and Schaefer, 1989, Brennan, 1990]. Clark [1996] characterizes this in terms of the concept of *closure*, or having good evidence that actions have succeeded. The collaborative rationality identified by Grice requires interlocutors to work for joint closure on any collaborative project, and particularly in talk exchange. Traum and Allen [1994] describe grounding in even stronger terms. They argue that speakers have an *obligation* to provide evidence of their understanding of other interlocutors and address the issues others raise, above and beyond the real-world collaborative interests they share.

Both kinds of approaches suggest an architecture for conversation where interlocutors regularly assess what they understand and estimate what their interlocutors understand. Many conversations are successful only if all understand one another. Accordingly, interlocutors bring goals, preferences or obligations for mutual understanding that they track on a par with the outcomes they track for sharing information and achieving real-world results.

The diversity of moves that accomplish grounding, as surveyed in Section 1.2.1, imposes further constraints on a theory of grounding. Grounding must, in Brennan's phrase [1990], be a matter of seeking and providing *evidence* for understanding. This evidence can take a variety of forms [see Clark, 1996, pp. 223ff]. It's clear that repetitions, as in (1.1), reformulations, as in (1.4), and assertions of understanding, as in (1.5), provide evidence about the addressee's level of understanding. But even followup utterances, as in (1.2), and assertions of non-understanding, as in (1.3), do the same. In fact, perhaps counterintuitively, followup utterances can provide quite good evidence about an addressee's understanding (or lack thereof), regardless of whether or not the addressee intends the utterance to do so. Conversely, frank assertions of understanding that reveal only the addressee's own judgment may be quite unreliable.

These cases show that interlocutors face extended periods of transient uncertainty during grounding. During these periods, they must assess the evidence they have about grounding, and trade off the costs of further clarification against the risks of persistent misunderstanding as they decide on their next contributions to the conversation [Horvitz and Paek, 2001]. The resulting dynamics of interaction can be approximated by models of grounding acts which insist that contributions must be acknowledged and accepted before they can be admitted into an incrementally-updated representation of common ground and before the conversation can move forward [Traum, 1994, Matheson et al., 2000]. But in the general case, interlocutors' reasoning must be more nuanced than such models might suggest.

Coherence theories of the organization of discourse offer complementary insights into models of grounding. Useful surveys are Hobbs et al. [1993], Kehler [2001] and Asher and Lascarides [2003]. All the moves we make in conversation respect the distinctive structure and process of human collaborative activity. Grounding moves are no exception. Aproaches to grounding can therefore benefit from detailed models of discourse interpretation, including both overarching general constraints and specific syntatic, semantic and interactive resources that are available for providing evidence of understanding.

Conversation is structured hierarchically, with short segments that address focused subtasks nested within longer segments that address larger tasks [Grosz and Sidner, 1986]. This structure accords with the difficulty we have when we approach tasks in the wrong order or have to revisit issues we thought were resolved. This is an important consideration in when and how to ground.

Another such consideration is the context-dependence of utterance interpretation. Many utterances, including those typically relevant for grounding, express a specific relationship to the information presented and the open questions raised in prior discourse. Many also involve implicit references to salient entities from the discourse context. Both kinds of contextual links must be resolved as part of recognizing the speaker's contribution with the utterance. It is largely through these links that utterances provide evidence about the speaker's acceptance and understanding of prior contributions to conversation [Lascarides and Asher, 2009,

Stone and Lascarides, 2010]. Thus, these links explain efficient utterances that ground implicitly. These links matter just as much in problematic cases, where interlocutors (and systems) must avoid utterances with contextual connections that might appear to ground.

Discourse theory also highlights the specific grammatical and interactive resources that make it easy for interlocutors to provide evidence of their understanding. For example, the rules of grammar sometimes interpret fragmentary utterances as if they occur directly in syntactic construction with previous utterances, as in the successive expansions of (1.3). See Gregoromichelaki et al. [2011] for a wide range of further cases. Other types of fragments, such as the reprise fragments of (1.1), seem to carry semantic constraints that closely tie their interpretation to those of prior utterances [Ginzberg and Cooper, 2004, Purver, 2004]. Both cases create interpretive connections that readily signal grounding.

We also have distinctive knowledge about how to negotiate as part of a collaborative activity. Collaborative negotiation typically involves a distinctive inventory of possible contributions to the activity: negotiators can make proposals, revise them, accept them conditionally or unconditionally, and so forth [Sidner, 1994, Carberry and Lambert, 1999, Eugenio et al., 2000]. Collaborative negotation is frequently modeled through additional layers of discourse structure, which explicitly represent the ordinary contributions of utterances as the object of meta-level discussion [Carberry and Lambert, 1999, Allen et al., 2002]. Models along these lines naturally describe interlocutors' collaborative efforts to agree on linguistic utterances and their interpretations [Heeman and Hirst, 1995].

1.2.3 Grounding as problem solving

The examples of Section 1.2.1 and the theoretical accounts of Section 1.2.2 portray grounding strategies as flexible responses to a speaker's information state and goals, given the affordances of grammar and collaboration. As a framework for realizing such responses in dialogue systems, we advocate a characterization of grounding as problem solving.

Problem solving is a general perspective on flexible intelligent behavior. See Newell [1982] or Russell and Norvig [1995]. A problem-solving system is endowed with general knowledge about the actions available to it and their possible effects, and with goals or preferences that it must strive to make true through the behavior it chooses. The system approaches a new situtaion by distinguishing key features of the situation: those that the system needs to change, on the one hand, and those that define the system's opportunities to act, on the other. This representation constitutes a problem for the system. The system solves such problems by reasoning creatively. It systematically explores possibilities for action and predicts their results, until it improvises a program of actioon that it can use to further its goals to a satisfactory degree in the current situation.

To treat grounding as problem solving is to design a conversational agent with knowledge about possible utterances that includes the contributions that

utterances can make and the evidence that utterances offer about their speaker's understanding. In the case of grounding, this knowledge must describe general models of collaborative discourse, along with particularly relevant grammatical constructions and negotiation moves. The conversational agent must then track its own understanding and the understanding of its interlocutors, and aim to reduce the uncertainty in these understandings to a satisfactory level. In particular, a specific pattern of ambiguity in a specific conversational situation is a trigger for synthesizing a new utterance whose interpretation highlights the potentailly problematic nature of the interaction and initiates a possible strategy to resolve it.

Methodologically, our appeal to problem solving plays two roles. First, it serves to link system design to empirical and theoretical results about grounding, by emphasizing the knowledge that is realized in systems rather than the specific algorithms or processing through which systems deploy that knowledge. Second, it provides a transparent explanation of the design of certain kinds of grounding systems: namely, those that navigate through a large and heterogeneous space of possible utterances to synthesize creative utterances for new situations. This suits our purpose in documenting and analyzing our COREF system in Section 1.3.

Problem solving is a theoretical perspective rather than a specific technique. Dialogue strategy is often engineered using specific models, such as Partially-observable Markov Decision Processes (POMDPs) [Williams and Young, 2007], which clarify specific aspects of the system's behavior. Mathematically, a POMDP starts from a precise probabilistic definition of what any utterance can contribute and the evidence a system gets about these contributions moment by moment. It derives an overall strategy that chooses utterances with an eye to long-term success. This accounts for the system's reasoning in maintaining uncertainty about the interaction and in making quantitative tradeoffs between gathering information and advancing task goals. By contrast, detailed issues of choice in natural language generation depend on methods that let us compute what new utterances can do, creatively, across an open-ended generative space. Where POMDP models simply assume that this computation can be accomplished straightforwardly, a problem-solving perspective lets us clarify the contributions of techniques for action representation, discourse modeling, and the effective management of search. Conversely, of course, the information a POMDP encodes also needs to be present in a problem-solving model to handle the decisions that are emphasized in POMDP research. This is why the choice of presentation is in part a matter of theoretical perspective.

Our approach to grounding elaborates the model of NLG as problem solving from the SPUD generator [Stone et al., 2003]. SPUD solves problems of contributing specific new information to interlocutors against the backdrop of a determinate common ground. SPUD's linguistic knowledge takes the form of a lexicalized grammar with entries characterized in syntactic, semantic and pragmatic terms. A model of interpretation predicts how an utterance with a given semantics can link up to the context to convey relevant information. A solution

to an NLG problem is a syntactically complete derivation tree whose meaning, as resolved unambiguously in context, contributes all the necessary information to the conversation, without suggesting anything false.

SPUD's problem-solving approach offers an integrated account of a range of generation effects, including the aggregation of related information into complex sentences, the planning of referring expressions, and the orchestration of lexical and syntactic choices. Importantly, it tracks the creative, open-ended ways in which these effects overlap in complex utterances. Moreover, the problem-solving approach emphasizes the role for declarative techniques in system design. The knowledge SPUD uses can be derived from diverse data sources, such as grammars designed for language understanding, linguistic analyses of target corpora, and machine learning over attested or desired uses of utterances in context.

SPUD's generation model needs to be changed in a number of ways to handle the kinds of grounding phenomena illustrated in Section 1.2.1. We need to handle the many different kinds of contributions that utterances can make to conversation, besides just providing new information. We need to predict not just the contributions that utterances make directly to the conversation, but also the indirect effects that utterances can have on interlocutors' information, especially their assessments of grounding. Finally, we need to take into account possible uncertainties in the context, as we calculate what interpretations an utterance could have or whether the addressee will understand it as intended. In our work with the COREF system, we have developed an initial approach to these extended models. In Section 1.3, we outline our approach and summarize its role in enabling our system to exhibit a rich new range of grounding strategies.

1.3 An NLG model for flexible grounding

COREF participates in a two-agent object identification game which we adapted from the experiments of Clark and Wilkes-Gibbs [1986] and Brennan and Clark [1996]. Our game plays out in a special-purpose graphical user interface, which can support either human–human or human–agent interactions. The objective is for the two agents to work together to create a specific configuration of objects, or a "scene", by adding objects into the scene one at a time. The two players participate from physically-separated locations so that communication can only occur through the interface. Each has their own version of the interface, which displays the same set of candidate objects but in differently-shuffled spatial locations. The shuffling undermines the use of spatial expressions such as "the object at the top left".[1]

[1] Note that in a human–human game, there are literally two versions of the graphical interface on separate computers. In a human-agent interaction, the agent uses a software interface that provides the same information that the graphical interface would provide to a human

As in the experiments of Clark and Wilkes-Gibbs [1986] and Brennan and Clark [1996], one of the players, who plays the role of *director*, instructs the other player, who plays the role of *matcher*, which object goes next. As the game proceeds, the next object is automatically determined by the interface and privately indicated to the director using a blue arrow. The director's job is then to get the matcher to click on (their version of) this object.

To achieve agreement about a target object, the two players can talk back and forth using written English, in an instant-messaging style modality. Each player's interface provides a real-time indication that their partner is "Active" while their partner is composing an utterance in their text box, but the interface does not show in real-time what characters are being typed. Thus, it is not possible for a player to view or interpret an utterance by their partner incrementally (as it is typed, word by word). Once the `Enter` key is pressed, the utterance appears on the screens of both players as part of a scrollable history field in the interface, which provides full access to all the previous utterances in the dialogue.

When the matcher clicks on an object they believe is the target, their version of that object is privately moved into their scene. The director has no visible indication that the matcher has done this. However, the director needs to click the `Continue (next object)` button in order to move the current target into the *director's* scene, and move on to the next target object. This means that the players need to discuss not just what the target object is, but also whether the matcher has added it into their scene, so that they can coordinate on the right moment to move on to the next object. If all the coordination succeeds perfectly, then after the director and matcher have completed a series of objects, they will have created the exact same scene in their separate interfaces.

We have run a series of evaluation experiments in which COREF interacts with human subjects [DeVault and Stone, 2007, DeVault, 2008]. In this paper, we illustrate qualitative aspects of COREF's grounding behavior by describing the reasoning that COREF uses to generate questions for the user in three key interactions from this data set. In (1.6), COREF plays the matcher, but begins by trying to guess the next object to add. With its question *is the target pink?*, COREF attempts to reduce its ignorance about the display.

(1.6) Agent: is the target pink?
a18: no
a18: it is brown
Agent: do you mean dark brown?
a18: yes

The display features light brown and dark brown objects, as well as pink ones. COREF's clarification question *do you mean dark brown* attempts to disambiguate which color characterizes the object.

participant. For example, the agent is not aware of the spatial locations of objects on the screen of its human partner.

In (1.7), COREF is the director.

(1.7) Agent: the blue circle
s1: ok
Agent: did you add it?
s1: yes

COREF interprets the user's *ok* as ambiguous. Maybe the user has added the blue circle to the display; maybe the user is simply signaling their understanding. COREF's *did you add it?* is a reminder designed to move the conversation forward in either case.

Finally, COREF as matcher in (1.8) finds two objects which might plausibly be described as *beige diamonds*.

(1.8) s2: the object is the beige diamond
Agent: do you mean the solid beige diamond?
s2: no
s2 : i mean the empty beige diamond
s2: did you add it?
Agent: done

COREF's clarification question distinguishes the two objects based on their fill.

A uniform problem-solving model leads to each of these questions in COREF. The model is distinguished by its flexible mapping from utterances to task actions (Section 1.3.1), its explicit representation of uncertainty and its effects on interpretation (Section 1.3.2), and its specification of generation problems in terms of a pattern of acceptable interpretations that move the conversation forward despite uncertainty (Section 1.3.3).

1.3.1 Utterances and contributions

Our modeling challenge in COREF's domain is to characterize the different contributions that utterances can make. Our approach is to use an architecture of *dialogue acts*. We follow Bunt [1994, 1996, 2000] in understanding a dialogue act as a "functional unit used by a speaker to change the context". In particular, each dialogue act comprises a *semantic content* and a *communicative function*. The semantic content is the information the speaker is introducing into the context; for example, some proposition p. The communicative function is the way in which that information has been inserted in the context in order to play its intended role; for example, with a role like INFORM, YN-QUESTION, or CORRECT [Bunt, 2000]. Together, the communicative function and semantic content determine an update function that maps the previous context to the new context that results from the dialogue act [Larsson and Traum, 2000].

Researchers commonly hypothesize dialogue acts that specifically govern grounding. For example, Bunt [1994] offers a top-level distinction between *dialogue control acts* and *task-oriented acts*, and then subdivides dialogue control

acts into *feedback acts*, *discourse structuring acts*, and *interaction management acts*. Meanwhile, Traum and Hinkelman [1992] distinguish four top-level types of *conversation acts*: *turn-taking*, *grounding*, *core speech acts*, and *argumentation*. The DAMSL annotation scheme [Core and Allen, 1997] captures grounding by distinguishing between the *backward* and *forward communicative functions* of *communicative acts*.

By contrast, we have designed COREF with a special-purpose set of about twenty action types. Our inventory endows COREF with detailed knowledge about problematic reference and other structured collaborations. Grounding is a side effect of interlocutors' reasoning and strategies in using these actions.

Our action set builds on stack–based models of clarification subdialogues [Ginzberg and Cooper, 2004, Purver, 2004], collaborative negotiation models [Sidner, 1994, Carberry and Lambert, 1999, Eugenio et al., 2000], and the use of collaborative discourse theory to characterize user interface tasks [Rich et al., 2001]. For example, one action type, `pushCollabRef`$[D, M, T]$ lets the director D initiate collaborative reference with matcher M to a target T. This action type is tacit: we identify its occurrence with the mental decision by D to collaboratively identify target T to M. Its update is to push a new task onto a stack of active tasks in the context and set up an initially empty constraint network recording the information that the interlocutors have used to describe T. Once this task is underway, the director can perform a dialogue act, `addcr`$[T, C]$, whose updates adds the additional constraint C to the constraint network.

The identification by the matcher that the target T is some entity R is captured by a tacit mental action `setVarValue`$[T, R]$. Its effect is to add the proposition `varValue`(T,R) to the context. After identifying R as a target object for the scene, the matcher can take the tacit action `addToScene`$[R]$. Although the matcher needs to carry out the action `ClickToAdd`$[R]$ to add the object, this is invisible to the director. The effect of `ClickToAdd`$[R]$ is to physically move the object into the scene part of the matcher's experiment interface. The update associated with agent A taking action `addToScene`$[R]$ is that the proposition `inScene`$[R,A]$ is added to the list of propositions in the context.

In addition, COREF's interpretation process causes certain additional updates to the dialogue state whenever an observable action occurs. These additional updates allow COREF to keep track of its uncertainty in interpretation; they also push a `ManageAmbiguity` subtask, which makes it coherent for the observer x of action A to perform follow-up actions to deal with uncertainty.

We associate utterances with these dialogue acts by modeling utterance interpretation as an intention-recognition problem. See Stone [2004] for theoretical motivation of this framework. We assume that each utterance by a speaker is intended to *generate* [Goldman, 1970, Pollack, 1986] one or more dialogue acts. In addition, we assume that utterances are often intended to reveal the occurrence of certain tacit dialogue acts indirectly, through a kind of conversational implicature we have called *enlightened update* [Thomason et al., 2006]. To recognize the speaker's intention requires integrating three sources of relevant information:

the logical structure of the ongoing activity in the conversation, the current state of the dialogue, and grammatical constraints.

Our model of task structure takes the form of a function $N(A, s)$ that captures the set of alternative actions that agent A can coherently perform next in dialogue state s. This function captures similar regularities as accounts of *adjacency pairs*, the use of *dialogue grammars* to distinguish coherent sequences of dialogue acts (or speech acts) from incoherent ones, the use of state machines with dialogue acts (or speech acts) attached to the state transitions, and with plan-based models of dialogue coherence; see [Cohen, 1997] for a review of work using these techniques. This function is populated in two ways in COREF. Some of its results are determined directly from the topmost task from the stack of active tasks in the dialogue. Handbuilt graphs outline all possible paths for deploying COREF actions to pursue any task. Other results list actions can happen coherently at a wide variety of points in a dialogue. These actions include operations that push subtasks, including describing the current state of the dialogue and resolving a yes–no or wh–question, and operations that manage flow through the dialogue, such as abandoning the topmost dialogue subtask.

COREF uses its model of coherent contributions as constraints on the possible intentions of interlocutors at any point in the conversation. The actual constraints are derived in two stages to account for the possibility of tacit actions signaled by implicature. Given a representation s that captures a possible state for the dialogue after the last overt action, COREF builds a representation, which we call a *horizon graph*, that captures the coherent ways the speaker might have tacitly intended to update the dialogue before making an overt contribution. For efficiency, we use handbuilt heuristics to discard tacit actions that *could* be performed coherently, according to COREF's task models, but which an interpreter would never actually hypothesize in this situation, given what we know about the overt action that the agent has chosen.

In the second step, COREF uses the horizon graph to solve any constraints associated with the observed action. This step instantiates any free parameters associated with the action to contextually relevant values. For utterances, the relevant constraints are identified using a grammar that determines both presupposed constraints—which must hold in the context—as well as schematic dialogue acts—which must be coherent in the context. The overall constraints associated with an utterance are determined by performing a bottom-up chart parse of the utterance, and joining the presuppositions and dialogue acts associated with each edge in the chart.

Our architecture of dialogue acts and interpretive inference allows us to reason about the contributions of question-asking moves in collaborative reference at a fine granularity. The questions of (1.6–1.8), for example, differ in whether the question is intended to add a constraint about a target referent—a property or object referenced in a previous utterance or the next thing to add to the scene—or whether it is intended to establish the occurrence of a specific event. The utterances also differ in the tacit moves that relate them to the ongoing discourse:

they may push a reminder task, begin a clarification subdialogue amplifying on a previous utterance, or simply introduce a specific issue whose resolution contributes to the current task. Each of these interpretations is obtained by the process we have sketched here: constraint-satisfaction inference that matches the grammatical structure of a candidate utterance against a context-specific inventory of coherent contributions to the conversation.

1.3.2 Modeling uncertainty in interpretation

The examples of (1.6–1.8) involve user utterances for which the interpretation models of Section 1.3.1 offer multiple plausible analyses. In (1.6) the presence of light brown and dark brown objects on the display translates into two possible ways that COREF could understand the user's reference to a color *brown*. In (1.7), the state of the task offers two different dialogue acts, each of which constitutes a possible resolution for *ok*. In (1.8), COREF sees two possible referents for *the beige diamond*. All the ambiguities are initially represented as multiple interpretations that assign coherent dialogue acts to the utterance in context.

When an interlocutor uses such ambiguous utterances, COREF models the evolving state of the conversation as uncertain. In particular, COREF tracks the dialogue acts possibly generated by the utterance, on each reading, by spawning a new thread of interpretation where those acts are assumed to occur. By spawning different threads of interpretation, capturing the user's alternative contributions, COREF can continue to assign coherent, principled interpretations to the user's ongoing utterances, and continue to participate in the dialogue.

COREF uses probabilistic reasoning to reallocate probability mass among the different threads of interpretation over time, while retaining a principled approach to linguistic interpretation in context within each thread. This updating allows COREF to model the clarification subdialogues in (1.6) and (1.8) as introducing and then resolving transient uncertainty. Thus in (1.6), the possible *light brown* interpretation creates a thread of interpretation in which the target has been constrained to have a light brown color, because of the user's contribution in making the utterance *it is brown* on its intended interpretation. When, however, the user answers *yes* to indicate that the intended color was dark brown rather than light brown, the answer is incompatible with this thread of interpretation. Accordingly, COREF assigns high probability to the other thread, which, COREF now realizes, has always tracked the correct understanding of the user's utterance *it is brown*. COREF's clarifications can resolve relevant ambiguities without necessarily pinpointing one correct thread of interpretation. In the reminder of (1.7), COREF ensures that the user clicks to add the next object to the scene. But because COREF recognizes that an effective reminder could prompt the user to add the object and *then* answer yes, COREF never discovers exactly what function the user originally intended for *ok*.

Since the present paper is focused on generation, we simply note that COREF's context for generation is often a set of alternative threads. We detail COREF's

interpretation and dialogue management under uncertainty, in particular our efforts to learn probability models from interacting with users, more fully in DeVault and Stone [2007] and DeVault and Stone [2009].

1.3.3 Generating under uncertainty

In COREF, the task of generation is to formulate an utterance that will make a clear positive contribution to the dialogue no matter what the true context turns out to be. COREF's utterances are represented along with specific links to the context and implicatures that spell out exactly how the utterance contributes to the ongoing task. COREF's uncertainty may therefore make it impossible to formulate an utterance with one definite interpretation.

For example, after *ok* in (1.7), our models predict that *did you add it?* will carry different implicatures depending on what the user intended to acknowledge. If the user meant only to show that they understood COREF's utterance, then *did you add it?* will implicate that COREF has given enough information so that the user should have identified the target. This implicature is not present if the user meant that they added the object to the scene—the user already knows and has acted on this. Accordingly, in addition to a set of dialogue acts to generate that serve as explicit communicative goals, COREF's generation problems are specified in terms of a set of *acceptable contributions* that represent reasonable ways to move the conversation forward. COREF aims for an utterance that is likely to achieve the specified communicative goals, and will make an acceptable contribution in any case.

COREF's model of context also shapes the problem of disambiguating its utterances. Disambiguation requires COREF to take into account both its own uncertainty and that of its interlocutor. On COREF's side, each of the possible states of the conversation represents a possible source of ambiguity. For example, if reference resolution predicts that an anaphoric expression would be resolved one way in one context, but another way in another context, COREF needs to represent and track the potential ambiguity. In addition, COREF must track the distinctive ambiguities that its interlocutor faces in identifying the implicatures that COREF is tacitly making. Recall that these implicatures are determined by a horizon graph that maps out coherent ways to continue the conversation.

COREF is designed to anticipate these ambiguities in a streamlined way during generation, without searching over implicatures and corresponding states of the conversation. COREF constructs a special model of constraint interpretation for generation, which we call a *pseudo-context*. A pseudo-context C behaves as follows. For each state s that represents one possible thread of the conversation, and each accessible state s_i that is on the horizon graph from s, any resolution of the utterance's presuppositions and dialogue acts in s_i also counts as resolution of the utterance's presuppositions and dialogue acts in C. The effect of this behavior is to increase the amount of ambiguity that is visible in generation beyond what would be present in a single context, to reflect ambiguities arising

throughout the possible horizon. To completely disambiguate its implicatures, the generator should distinguish an intended horizon state in the possible horizon graph [Thomason et al., 2006], so it would perhaps be ideal to use this reasoning explicitly in generation. However, we have not yet assessed how important this reasoning could be to COREF's behavior in practice.

In COREF, generation occurs in two steps. The first phase starts from a set of communicative goals and a set of acceptable contributions, and a pseudo-context for resolving meaning. The communicative goals are selected by the dialogue manager, targeting both the system's uncertainty and the specific dialogue state that COREF thinks is most likely. The set of acceptable contributions is determined by COREF's handbuilt dialogue policy. The pseudo-context, as just described, outlines possible ways the presuppositions and schematic dialogue acts could be resolved given the system's uncertainty and implicatures about the state of the conversation. COREF searches through its lexicalized grammar, extending a provisional utterance, word-by-word, until it finds a complete derivation that achieves the specified communicative goals unambiguously in the extended pseudo-context. This is essentially the same search algorithm as SPUD, but using a richer model of interpretation.

The second phase applies once the generator produces a candidate utterance. COREF interprets the utterance as though it were actually uttered. (This process evaluates the utterance on a context-by-context basis; no pseudo-context is involved. This step therefore offers an accurate assessment of any remaining implicatures and ambiguities in the utterance.) If all of the actions that appear in all the interpretations found are acceptable to COREF, then COREF accepts the output utterance. 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*. This scenario is one in which COREF makes an underspecified contribution.

If one or more of the interpretations for an output utterance is unacceptable to COREF, it reconsiders its dialogue policy by formulating a different communicative goal or different implicatures. This process repeats until an acceptable utterance is found, or until all communicative options are exhausted.

1.3.4 Examples

Across its conversations with our human subjects, COREF asked users 559 questions of 90 distinct sentence types. Examples (1.6–1.8) illustrate the different questioning moves COREF used. The variation in COREF's utterances come in how COREF constructs specific questions to achieve specific dialogue moves in specific contexts. For example, COREF uses different vocabulary to describe properties in the display depending on the salient contrasts, uses different referring expressions to describe objects depending on the alternative objects, and varies in its uses of full referring expressions versus pronouns depending on the dialogue history. Analyses of the user logs reveals that these questions are usually

effective in moving the conversation forward, for example by resolving COREF's uncertainty in cases of ambiguity. See DeVault [2008] for full details about these experimental results.

As we have seen, the generation model in COREF differs from its closest relative, SPUD, in several ways. COREF models a full range of dialogue acts, communicated both explicitly and by implicature. COREF tracks uncertainty about the context, and is correspondingly more flexible in how it assesses utterance interpretation and what counts as a successful output for generation. To show how these innovations are crucial to grounding in COREF, we consider in detail the reasoning that COREF uses to produce *did you add it?* in (1.7).

COREF starts from an uncertain context. There are two possible threads of interpretation in the dialogue: one where the user has just acknowledged COREF's description (this is the most probable in COREF's model of interpretation), and another where the user has just asserted that they have put the next target in place. COREF sets its communicative goal based on the coherent options in the most probable context. We briefly survey these options, as determined by COREF's horizon graph. Even though COREF is uncertain what the user's utterance of *ok* meant, COREF is willing to tacitly decline the opportunity to clarify — perhaps it could do so by uttering something like *what do you mean ok?* — by performing a `tacipNop` action. Moving forward, COREF's rules for action acceptability allow it to implicate, on the grounds that COREF's utterance *the blue circle* added a constraint that uniquely identified the target object, that the user must have identified the target object. COREF's domain model captures the user's mental identification of the target object as action `s1 : setVarValue[t3, e2_2]`. Here `s1` is COREF's symbol for the user, `t3` is COREF's symbol for the target of the current collaborative reference episode, and `e2_2` is COREF's symbol for the next object to add to the scene. The effect of this tacit mental action is to make it part of the context that target `t3` refers to object `e2_2`. COREF's acceptability rules further allow the agent to push a `Remind` task, which allows COREF to coherently ask whether s1 has added the object to their scene. COREF's policy is to try to ask for clarification whenever the state is ambiguous, so COREF's dialogue manager focuses on this state in the horizon graph and proposes to generate the dialogue act `COREF : askYNQ[past[s1, addToScene[e2_2]]]`. The dialogue move offers a yes-no question to establish whether the user has added this object to the scene.

In this situation, in order to proceed with generation, COREF anticipates interpretations in both states by constructing an expanded pseudo-state that maps out a range of possibilities. These include the dialogue states just surveyed, as well as corresponding dialogue states on the horizon graph for the other thread of interpretation. Search through the grammar identifies *did you add it?* as an unambiguous utterance expressing the needed dialogue act in this pseudo-state.

COREF then explicitly assesses how the user will interpret an utterance of *did you add it?* This involves recognizing the specific implicatures the utterance will generate in context, using the the horizon graph. COREF predicts that *did*

you add it? supports two interpretations, one from each of the current possible dialogue states. If the user's previous *ok* simply contributed a `nop`, the utterance supports the interpretation in (1.9):

(1.9) ⟨`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)]`⟩

This is the specific set of contributions that COREF makes with this utterance in this specific context: COREF is declining to manage any perceived uncertainty regarding s1's utterance of *ok,* implicating that `s1` has identified the target object, implicating the start of a reminder subtask task, overtly asking a yes–no question as part of that reminder task, and putting object `e2_2` into focus.

If the user's previous *ok* contributed more, COREF associates *did you add it?* with the interpretation in (1.10):

(1.10) ⟨`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)]`⟩

This interpretation is similar, except now COREF would not be implicating that the user has identified the target object.

Because COREF finds both of these contributions acceptable, in the corresponding dialogue states, COREF's generator returns action sequence ⟨`say`[`did you add it`?]⟩, along with the anticipated user interpretations. We view COREF as working collaboratively to move the conversation forward, but making an underspecified contribution. If the user intended to acknowledge COREF's description, then COREF's contribution *is* (1.9). If the user intended to assert that the target was in place, then COREF's contribution *is* (1.10). COREF accepts the utterance because it is willing to be recognized as making either contribution. In this way, COREF models itself as having extended two threads of coherent interpretation which it spawned upon perceiving the ambiguity in the user's utterance of *ok*.

1.4 Alternative approaches

Our work showcases the possibility of achieving grounding in conversational systems through general problem solving. In Section 1.5, we argue that such approaches will be particularly important for embodied conversational agents, in emerging applications such as human–robot interaction. In these domains, many

different kinds of actions can give evidence about interlocutors' understanding of one another, and a problem-solving model may be necessary to capture the complexity and diversity of the reasoning involved.

Do not underestimate the difficulty of realizing general problem solving in a working system, however. A problem-solving system needs to reason from general and accurate knowledge, which can be extremely challenging to specify by hand or learn from data. A problem-solving system also needs powerful search mechanisms to explore the consequences of its knowledge and synthesize solutions; these mechanisms usually require a combination of careful engineering and substantial computational resources. Moreover, the inference that problem-solving systems do makes their results unpredictable, exacerbates the errors associated with incorrect knowledge, and makes it difficult to offer guarantees about system behavior and performance. COREF has a number of limitations of this sort. One unfortunate interaction between COREF's construction of pseudo-contexts and its characterization of acceptable interpretations led to its occasional use of *do you mean it?* as a clarification question.

For all these reasons, applied spoken dialogue systems typically do not plan grounding utterances with general problem-solving models. Alternative implementation frameworks can lead to insights on important aspects in dialogue strategy where COREF is weak, such as empirical models of understanding, robust handling of uncertainty, and the ability to make quantitative tradeoffs in generation. We review some of this research in this section. Of particular note are abstract models of dialogue in terms of qualitative grounding moves, which minimize the need for reasoning about uncertainty in conversation (Section 1.4.1); models of user state that focus probabilistic inference to restrict the need for open-ended problem solving for generation under uncertainty (Section 1.4.2); and feature engineering to avoid the need for deep generalizations about how systems should address their own or their users' uncertainty (Section 1.4.3).

1.4.1 Idealizing Incremental Common Ground

Historically, linguists and philosophers have characterized state in conversation qualitatively in terms of a notion of mutual knowledge or common ground [Stalnaker, 1974, 1978, Clark and Marshall, 1981]. This is a body of information that interlocutors know they can rely on in formulating utterances. A simple way to adapt natural language generation techniques to dialogue, and handle grounding, is through rules that approximate the incremental evolution of the common ground in conversation.

Purver [2004] offers a particularly influential model of common ground in problematic dialogue. He proposes that utterances, by default, update the common ground to reflect their contributions. However, utterances like clarification questions, which show that previous discourse may not have been understood well enough for the purposes of the conversation, trigger "downdates" that erase previous contributions from the common ground. (Earlier models by Traum [1994]

and Matheson et al. [2000] assign contributions a "pending" status until they are acknowledged and accepted by all interlocutors.)

Incremental common ground models are limited in their abilities to assimilate information that accrues across multiple utterances, particularly when misunderstandings are discovered late, as in (1.2). Such cases call for probabilistic models and Bayesian reasoning [DeVault and Stone, 2006, Stone and Lascarides, 2010]. However, incremental common ground models enable elegant natural language generation implementations, because they can describe communicative context and communicative goals in familiar qualitative terms. Stent [2002] describes in detail how grounding acts can be realized as fluid utterances in incremental common ground models.

1.4.2 Focusing Probabilistic Inference

A different direction in grounding research focuses on the accurate representation of the system's uncertain state. In noisy settings, including most speech applications, effective dialogue management depends on using all the available evidence about the state of the conversation. Researchers in spoken dialogue systems have gone a long way towards capturing the evidence about user utterances that's actually provided by speech recognition systems, and understanding the strategies that let systems exploit this evidence to maximize their understanding [Roy et al., 2000, Horvitz and Paek, 2001, Williams and Young, 2006, 2007].

This research adopts decision-theoretic frameworks for optimizing choice in the face of uncertainty, such as the POMDPs briefly reviewed in Section 1.2.3. Making these frameworks practical requires streamlining models of the conversation, so researchers typically focus on accurately modeling user's private state or goal. For example, in the human–robot dialogue application of Roy et al. [2000], the POMDP model captures uncertainty about the command the user is trying to give the robot. In the POMDP-based call center application of Williams [2008], the state is the directory listing (name and listing type) that the caller wishes to be connected to. Richer models of conversational state can be accommodated only with shortcuts in solution strategies. In particular, as we will see in Section 1.4.3, most work on decision-theoretic optimization of dialogue management does not actually represent the system's moment-to-moment uncertainty about the state of the conversation, and so learns to ask clarification questions indirectly.

Conversely, in selecting a system utterance, optimized dialogue systems generally draw from a small fixed inventory of utterances that have been hand-authored by a system designer to effectively convey specific messages. For grounding, for example, the system might select from general prompts, acknowledgments, explicit confirmation questions, and implicit confirmations that verbalize the system's understanding of a prior utterance (word-for-word) while moving forward in the dialogue. Optimization allows the system to tune its strategy and complete tasks for users even in the face of frequent errors in speech recognition.

Similar optimization techniques can orchestrate the choice of high-level generation strategies in dialogue [Lemon, 2011]. However, with only coarse tools for describing utterance structure and dynamics means, these frameworks offer limited insight into the the various kinds of coordinated, context-sensitive linguistic expressions that can potentially realize grounding actions.

1.4.3 Correlating Conversational Success with Grounding-Related Features

A different approach to dialogue management focuses on optimizing tradeoffs about how to ground directly based on dialogue outcome. This approach simplifies the dialogue model to include only directly observable features in the represented dialogue state, using Markov Decision Processes [Levin and Pieraccini, 1997, Levin et al., 1998]. Rather than represent the user's utterance as a hidden variable, the model simply represents the recognition result as actually delivered by the speech recognizer. The model captures the possibility that this result is incorrect probabilistically: the system sometimes gets a bad outcome in dialogue when it acts on what it has heard. Rather than reason explicitly about the effects of clarification on the system's information, the model is instead designed to distinguish those dialogue states in which the system has obtained confirmation of a recognized parameter and those in which it has not. The usefulness of grounding actions is revealed by differences in rates of dialogue failure across these different kinds of states.

These approaches are very appealing because they support powerful empirical methods for estimating model parameters and powerful computational techniques for optimizing dialogue strategies. The optimization metric can be tuned to the factors that actually govern dialogue success in specific applications [Walker, 2000]. Models of dialogue can be learned accurately from small amounts of dialogue data using bootstrapping and simulation [Rieser and Lemon, 2011]. And function-approximation and state-abstraction techniques make it possible to compute good strategies for complex dialogue models [Henderson et al., 2008]. However, because the models ultimately describe the effects of system decisions directly through observable features of the dialogue, all grounding depends on the insights of the system-builders in describing the dialogue state through the right features. For example, Tetreault and Litman [2006] explicitly study which features to include in a state representation based on the impact those features have on learned dialogue policies.

1.5 Future Challenges

One way to think of the streamlined grounding models that characterize many current systems is that they provide empirically-based protocols that succeed in practice in overcoming specific kinds of communication problems. These protocols work because they help to align a system's behavior with the true state of

the user. Problem-solving models, by contrast, aim to endow the system more open-ended abilities to explicitly reason about understanding. We believe that such abilities will become increasingly important as dialogue systems begin to handle richer interactions.

1.5.1 Grounding with Multimodal Communicative Action

One important direction for richer dialogue is the design of embodied conversational agents [Cassell, 2000] which contribute to conversation using the affordances of a physical robot or graphical character, including gestures, gaze, facial expressions, and modulations of position and posture of the body as a whole. By using complex communicative actions that pair spoken utterances with actions in other modalities, embodied conversational agents can communicate their understanding in flexible and natural new ways.

For example, Nakano et al. [2003] describe an embodied conversational agent that gives directions with reference to a tabletop map. The system detects and adapts to the nonverbal grounding cues that followers spontaneously provide in human–human conversations. For example, Nakano et al. [2003] found that when listeners could follow instructions, they would nod and continue to direct their attention to the map, but when something was unclear, listeners would gaze up toward the direction-giver and wait attentively for clarification. The system interpreted these cues as grounding actions using an incremental common ground model, and varied its generation strategies accordingly. Nakano et al. [2003] found that their system elicited from its interlocutors many of the same qualitative dynamics they found in human–human conversations.

Multimodal grounding requires a framework for generating and interpreting multimodal communicative acts. A common strategy is "fission", or distributing a planned set of dialogue acts across different modalities to match patterns that are derived from analyses of effective interactions in a specific domain. SmartKom [Wahlster et al., 2001] is an influential early example of multimodal fission for dialogue generation. By contrast, problem-solving models of multimodal generation, such as Cassell et al. [2000], Kopp et al. [2004], reason about the affordances and interdependencies of body and speech to creatively explore the space of possible multimodal utterances and synthesize utterances that link specific behaviors to specific functions opportunistically and flexibly.

The potential advantage of a problem-solving model is the ability to reason indirectly about grounding. This is important if, as theoretical analyses suggest [Lascarides and Stone, 2009], gesture parallels language in its context-sensitivity and discourse coherence, and so affords similar indirect evidence about grounding. Lascarides and Stone [2009] analyze a fragment of conversation in which one speaker explains Newton's Law of Gravity to his interlocutor. The speaker explains the logic of the equation in part through a series of gestures that depict the Galilean experiment of dropping a series of weights in tandem. His addressee demonstrates that she understands his explanation by gesturing her own Galilean

experiment, which tracks and eventually anticipates the speaker's own. The evidence the gesture gives of understanding, just like the evidence spoken words give in cases like (1.4), is inseparable from the interpretation the gesture gets by linking up with the context and contributing content to it.

Researchers have not yet attempted to link problem-solving models of grounding, such as those explored in Section 1.3, with sophisticated descriptions of the interpretation of the form and meaning of multimodal utterances. We believe that doing so will lead to a much broader and more natural range of grounding functions for nonverbal behaviors in embodied conversational agents.

1.5.2 Cognitive Constraints

Explicit communicative contributions aren't the only evidence embodied agents give of their understanding. When virtual humans [Swartout et al., 2006] realize their embodied behaviors through computational architectures that limit attention, focus information processing and trigger emotional responses in human-like ways, one side-effect of these cognitive mechanisms can be to make aspects of the agent's conversational strategies and judgments more legible to human interlocutors. This is another indirect kind of grounding.

When people speak in problematic situations, their utterances reveal their uncertainty in recognizable ways [Brennan and Willams, 1995, Swerts and Krahmer, 2005, Stone and Oh, 2008]. People are simply slower to respond in cases of uncertainty. They also make different appraisals of the process of the conversation and their contributions to it when they are uncertain. Uncertainty may feel difficult, for example, or may lead to a contribution that feels unsatisfactory. These appraisals shape interlocutors' affect and so influence their facial expressions. To the extent that virtual humans exhibit the same cognitive and affective dynamics, their uncertainty will also be recognizable. See Stone and Oh [2008] for a case study.

In fact, Sengers [1999] has argued that agents that aim to be understood must not only exhibit the right cognitive and affective dynamics—they must work actively to reveal these dynamics to their audience. Sengers focused on clarifying the goals and decisions of animated characters, by dramatizing how they see, react to and engage with events in their virtual environments. It is an open problem to integrate these techniques with approaches to grounding based on the analysis of communicative action. A problem-solving approach offers a natural and attractive strategy to do this, since it promises to describe the evidence that agents provide about understanding through their communication in probabilistic terms that are compatible with other evidence that might come from agents' attention, processing or emotion.

1.5.3 Task Action

A final source of evidence about grounding comes from the real-world activity that accompanies dialogue in task-oriented interactions. Understanding what teammates have done is continuous with reasoning about what they have said and what they have understood. For example, suppose a speaker has given an instruction to an addressee. The action the addressee performs to carry out the instruction gives very good evidence about how the addressee understood the speaker and what the addressee thought was expected. Thus, it's natural to associate instrumental actions with grounding functions, especially in settings such as human–robot interaction where conversation is used to coordinate embodied activities among physically copresent interlocutors.

A starting point for modeling such functions might be to model task actions as generating grounding acts in certain cases. For example, carrying out an expected action might constitute a specific form of acknowledgement. Such models could encode useful strategies for carrying out efficient dialogues that avoid misunderstandings. However, the study of human–human dialogue again suggests that more general problem-solving models will be necessary to reproduce people's collaborative use of physical action in grounding.

For example, consider the results of Clark and Krych [2004]. They analyzed dyadic conversations in which a director leads a matcher through the assembly of a Lego structure. As you might expect, directors' instructions often include installments, as in (1.1) and expansions, as in (1.3), formulated in real-time in response to feedback from the matcher. This feedback, however, is often action: Matchers pose blocks tentatively and use other strategies, including accompanying verbal and non-verbal communicative action, to mark their actions as provisional. Some of the most effective teamwork features tight coordination where directors offer short fragments to repeatedly correct proposed matcher actions. Clark and Krych [2004] cite one example where the director's iterative critique of four successive poses—over just four seconds—frees the interlocutors from having to agree on an precise description of the complex spatial configuration of a difficult-to-describe piece.

COREF's problem-solving model already interprets observed task actions using the same intention-recognition framework as it uses to interpret utterances. That means that COREF expects these actions to meet the constraints established by the interlocutors in prior conversation as well as the natural organization of the ongoing activity. Thus, as with an utterance, COREF can simultaneously enrich its understanding of the action and resolve uncertainty in the context by reconciling its observations and its interpretive constraints.

COREF, however, is a long way from the fluidity found in human–human dialogues like Clark and Krych's [2004]. We suspect that modeling the grounding function of provisional or even incorrect actions requires extending the kinds of models in COREF with a generative account of the relationship between actions and instructions that factors in the underspecification, ambiguities and

errors common with natural language descriptions. See Stone and Lascarides [2010]. The negotiation involved, meanwhile, requires richer accounts of the use of fragmentary utterances to link up with and coordinate ongoing real-world action. These phenomena again highlight the long-term research opportunities and challenges of problem-solving models of grounding.

1.6 Conclusion

A fundamental problem in deploying natural language generation in dialogue systems involves enriching models of language use. Dialogue systems need to be able handle problematic interactions, and that means that they cannot simply exploit static models of form and meaning. They need to be able to negotiate their contributions to conversation, flexibly and creatively, despite missteps and uncertainties, across extended interactions. Natural language generation, as a field, is just beginning to engage meaningfully with these requirements and the challenges they bring. The most important problems are open, and this chapter has given a corresponding emphasis to exploratory and open-ended research.

In particular, our focus has been on problem-solving models—general accounts that help us operationalize the reasoning for synthesizing productive contributions in problematic situations. Problem solving provides a methodological tool for systematizing the grounding behaviors that people use in human–human conversation, for understanding the knowledge of activity, context and language that underpins grounding behaviors, and for mapping the possible interactions between language, embodied communication and cognition, and task action in keeping conversation on track. Of course, we can build on systematic thinking about grounding in dialogue in a wide range of implementations. Simpler and more constrained frameworks often provide the most efficient and robust realization of the insights of more general models.

Indeed, for the moment, problem-solving models may be most important as a bridge between descriptive accounts of human–human conversation and the strategies we choose to realize in practical dialogue systems. Descriptive analyses do surprisingly well by characterizing grounding in common-sense terms: people in problematic dialogues offer evidence of understanding, act collaboratively, negotiate, reach agreement. Computational models—and their limits—remind us how subtle and sophisticated this common-sense talk really is. The concepts involved tap directly into our most powerful principles of social cognition, principles for which science offers only the barest sketch. Linguistic meaning ranks among the triumphs of our abilities to relate to one another. In understanding, systematizing and implementing grounding in conversational agents, we deepen and transform our understanding of those abilities, and the abilities themselves.

References

Allen, J. F., Blaylock, N., and Ferguson, G. (2002). A problem solving model for collaborative agents. In *The First International Joint Conference on Autonomous Agents & Multiagent Systems, AAMAS 2002, July 15-19, 2002, Bologna, Italy, Proceedings*, pages 774–781.

Asher, N. and Lascarides, A. (2003). *Logics of Conversation*. Cambridge.

Brennan, S. E. (1990). *Seeking and Providing Evidence for Mutual Understanding*. PhD thesis, Stanford University.

Brennan, S. E. and Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology*, 22(6):1482–1493.

Brennan, S. E. and Willams, M. (1995). The feeling of another's knowing: prosody and filled pauses as cues to listeners about the metacognitive states of speakers. *Journal of Memory and Language*, 34:383–398.

Bunt, H. (1996). Interaction management functions and context representation requirements. In LuperFoy, S., Nijholt, A., and van Zanten, G. V., editors, *Dialogue Management in Natural Language Systems. Proc. of 11th Twente Workshop on Language Technology, University of Twente, Enschede*, pages 187–198.

Bunt, H. (2000). Dialogue pragmatics and context specification. In Bunt, H. and Black, W., editors, *Abduction, Belief and Context in Dialogue. Studies in Computational Pragmatics*, pages 81–150. Amsterdam: Benjamins.

Bunt, H. C. (1994). Context and dialogue control. *THINK Quarterly*, 3:19–31.

Carberry, S. and Lambert, L. (1999). A process model for recognizing communicative acts and modeling negotiation subdialogues. *Computational Linguistics*, 25:1–53.

Cassell, J. (2000). Embodied conversational interface agents. *Communications of the ACM*, 43(4):70–78.

Cassell, J., Stone, M., and Yan, H. (2000). Coordination and context-dependence in the generation of embodied conversation. In *Proceedings of INLG*.

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

Clark, H. H. (1993). *Arenas of Language Use*. University of Chicago.

Clark, H. H. and Krych, M. (2004). Speaking while monitoring addressees for understanding. *Journal of Memory and Language*, 50:6281.

Clark, H. H. and Marshall, C. R. (1981). Definite reference and mutual knowledge. In Joshi, A., Webber, B., and Sag, I., editors, *Elements of Discourse*

Understanding, pages 10–63. Cambridge University Press, Cambridge, England.

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

Clark, H. H. and Wilkes-Gibbs, D. (1986). Referring as a collaborative process. In Cohen, P. R., Morgan, J., and Pollack, M. E., editors, *Intentions in Communication*, pages 463–493. MIT Press, Cambridge, Massachusetts, 1990.

Cohen, P. (1997). Dialogue modeling. In Cole, R., Mariani, J., Uszkoreit, H., Varile, G. B., Zaenen, A., and Zampolli, A., editors, *Survey of the State of the Art in Human Language Technology (Studies in Natural Language Processing)*, pages 204–210. Cambridge University Press.

Core, M. G. and Allen, J. F. (1997). Coding dialogues with the DAMSL annotation scheme. In Traum, D., 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, D. (2008). *Contribution Tracking: Participating in Task-Oriented Dialogue under Uncertainty*. PhD thesis, Department of Computer Science, Rutgers, The State University of New Jersey, New Brunswick, NJ.

DeVault, D., Kariaeva, N., Kothari, A., Oved, I., and Stone, M. (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, D. and Stone, M. (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, D. and Stone, M. (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, D. and Stone, M. (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.

Eugenio, B. D., Jordan, P. W., Thomason, R. H., and Moore, J. D. (2000). The agreement process: An empirical investigation of human-human computer-mediated collaborative dialogue. *International Journal of Human-Computer Studies*, 53:1017–1076.

Furnas, G. W., Landauer, T. K., Gomez, L. M., and Dumais, S. T. (1987). The vocabulary problem in human-system communications. *Communications of the ACM*, 30:964–971.

Ginzberg, J. and Cooper, R. (2004). Clarification, ellipsis and the nature of contextual updates in dialogue. *Linguistics and Philosophy*, 27(3):297–365.

Goldman, A. (1970). *A Theory of Human Action*. Prentice-Hall.

Gregoromichelaki, E., Kempson, R., Purver, M., Mills, G. J., Cann, R., Meyer-Viol, W., and Healey, P. G. T. (2011). Incrementality and intention-recognition in utterance processing. *Dialogue and Discourse*, 2(1):199–233.

Grice, H. P. (1975). Logic and conversation. In Cole, P. and Morgan, J., editors, *Syntax and Semantics III: Speech Acts*, pages 41–58. Academic Press, New York.

Grosz, B. J. and Sidner, C. L. (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Heeman, P. A. and Hirst, G. (1995). Collaborating on referring expressions. *Computational Linguistics*, 21(3):351–383.

Henderson, J., Lemon, O., and Georgila, K. (2008). Hybrid reinforcement/supervised learning of dialogue policies from fixed data sets. *Computational Linguistics*, 34(4):487–511.

Hobbs, J. R., Stickel, M., Appelt, D., and Martin, P. (1993). Interpretation as abduction. *Artificial Intelligence*, 63:69–142.

Horvitz, E. and Paek, T. (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.

Kehler, A. (2001). *Coherence, Reference and the Theory of Grammar*. CSLI.

Kopp, S., Tepper, P., and Cassell, J. (2004). Towards integrated microplanning of language and iconic gesture for multimodal output. In *Proceedings of the International Conference on Multimodal Interfaces (ICMI 2004)*, pages 97–104.

Larsson, S. and Traum, D. (2000). Information state and dialogue management in the TRINDI dialogue move engine toolkit. *Natural Language Engineering*, 6:323–340.

Lascarides, A. and Asher, N. (2009). Agreement, disputes and commitments in dialogue. *Journal of Semantics*, 26(2):109–158.

Lascarides, A. and Stone, M. (2009). Discourse coherence and gesture interpretation. *Gesture*, 9(2):147–180.

Lemon, O. (2011). Learning what to say and how to say it: Joint optimisation of spoken dialogue management and natural language generation. *Computer Speech & Language*, 25(2):210–221.

Levin, E. and Pieraccini, R. (1997). A stochastic model of computer-human interaction for learning dialogue strategies. In *Proceedings of Eurospeech*, pages 1883–1886, Rhodes, Greece.

Levin, E., Pieraccini, R., and Eckert, W. (1998). Using markov decision process for learning dialogue strategies. In *Proceedings International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Matheson, C., Poesio, M., and Traum, D. (2000). Modelling grounding and discourse obligations using update rules. In *Proceedings of NAACL*.

Nakano, Y. I., Reinstein, G., Stocky, T., and Cassell, J. (2003). Towards a model of face-to-face grounding. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL 2003)*, pages 553–561.

Newell, A. (1982). The knowledge level. *Artificial Intelligence*, 18:87–127.

Pollack, M. (1986). A model of plan inference that distinguishes between the beliefs of actors and observers. In Biermann, A. W., editor, *Proceedings of the 24th Meeting of the Association for Computational Linguistics (ACL)*, pages 207–215, Morristown, New Jersey. Association for Computational Linguistics.

Purver, M. (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, C., Sidner, C. L., and Lesh, N. (2001). Collagen: Applying collaborative discourse theory to human-computer interaction. *Artificial Intelligence Magazine*, 22(4):15–25.

Rieser, V. and Lemon, O. (2011). Learning and evaluation of dialogue strategies for new applications: Empirical methods for optimization from small data sets. *Computational Linguistics*, 37(1):153–196.

Roy, N., Pineau, J., and Thrun, S. (2000). Spoken dialog management for robots. In *The Proceedings of the Association for Computational Linguistics*.

Russell, S. and Norvig, P. (1995). *Artificial Intelligence: A Modern Approach*. Prentice Hall.

Sengers, P. (1999). Designing comprehensible agents. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI 1999)*, pages 1227–1232.

Sidner, C. L. (1994). Negotiation in collaborative activity: a discourse analysis. *Knowledge Based Systems*, 7(4):265–267.

Stalnaker, R. (1974). Pragmatic presuppositions. In Stalnaker, R., editor, *Context and Content*, pages 47–62. Oxford, New York, New York.

Stalnaker, R. (1978). Assertion. In Cole, P., editor, *Syntax and Semantics 9*. Academic Press, New York, New York.

Stent, A. J. (2002). A conversation acts model for generating spoken dialogue contributions. *Computer Speech and Language*, 16:313–352.

Stone, M. (2004). Communicative intentions and conversational processes in human-human and human-computer dialogue. In Trueswell and Tanenhaus, editors, *World-Situated Language Use*. MIT.

Stone, M., Doran, C., Webber, B., Bleam, T., and Palmer, M. (2003). Microplanning with communicative intentions: the spud system. *Computational Intelligence*, 19(4):314–381.

Stone, M. and Lascarides, A. (2010). Coherence and rationality in dialogue. In *Proceedings of the 14th SEMDIAL Workshop on the Semantics and Pragmatics of Dialogue*, pages 51–58, Poznan.

Stone, M. and Oh, I. (2008). Modeling facial expression of uncertainty in conversational animation. In Wachsmuth, I. and Knoblich, G., editors, *Modeling Communication with Robots and Virtual Humans*, pages 57–76. Springer.

Swartout, W., Gratch, J., Hill, R. W., Hovy, E., Marsella, S., Rickel, J., and Traum, D. (2006). Toward virtual humans. *AI Mag.*, 27(2):96–108.

Swerts, M. and Krahmer, E. (2005). Audiovisual prosody and feeling of knowing. *Journal of Memory and Language*, 53(1):81–94.

Tetreault, J. and Litman, D. (2006). Using reinforcement learning to build a better model of dialogue state. In *Proceedings of the 11th Conference of the European Association for Computational Linguistics (EACL)*.

Thomason, R. H., Stone, M., and DeVault, D. (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, D. 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, D. R. and Allen, J. F. (1994). Discourse obligations in dialogue processing. In Pustejovsky, J., editor, *Proceedings of the Thirty-Second Meeting of the Association for Computational Linguistics*, pages 1–8, San Francisco. Association for Computational Linguistics, Morgan Kaufmann.

Traum, D. R. and Hinkelman, E. A. (1992). Conversation acts in task-oriented spoken dialogue. *Computational Intelligence*, 8(3):575–599.

Wahlster, W., Reithinger, N., and Blocher, A. (2001). SmartKom: multimodal communication with a life-like character. In *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH 2001)*, volume 3, pages 1547–1550.

Walker, M. A. (2000). An application of reinforcement learning to dialogue strategy selection in a spoken dialogue system for email. *Journal of Artificial Intelligence Research*, 12:387–416.

Williams, J. and Young, S. (2006). Scaling pomdps for dialog management with composite summary point-based value iteration (cspbvi). In *Proceedings AAAI Workshop on Statistical and Empirical Approaches for Spoken Dialogue Systems*.

Williams, J. and Young, S. (2007). Partially observable markov decision processes for spoken dialog systems. *Computer Speech and Language*, 21(2):393–422.

Williams, J. D. (2008). Demonstration of a pomdp voice dialer. In *Proc Demonstration Session, Annual Meeting of the Association for Computational Linguistics (ACL) with Human Language Technology Conference (HLT)*.