# Abductive Explanation of Dialogue Misunderstandings

**Susan McRoy and Graeme Hirst**
Department of Computer Science
University of Toronto
Toronto, Canada M5S 1A4

## Abstract

To respond to an utterance, a listener must interpret what others have said and why they have said it. Misunderstandings occur when agents differ in their beliefs about what has been said or why. Our work combines intentional and social accounts of discourse, unifying theories of speech act production, interpretation, and the repair of misunderstandings. A unified theory has been developed by characterizing the generation of utterances as default reasoning and using abduction to characterize interpretation and repair.

## 1 Introduction

When agents participate in a dialogue, they bring to it different beliefs and goals. These differences can lead them to make different assumptions about one another's actions, construct different interpretations of discourse objects, or produce utterances that are either too specific or too vague for others to interpret as intended. As a result, agents may fail to understand some part of the dialogue—or unknowingly diverge in their understanding of it—making a breakdown in communication likely. One strategy an agent might use to address the problem of breakdowns is to try to circumvent them, for example, by trying to identify and correct apparent confusions about objects or concepts mentioned in the discourse [Goodman, 1985; McCoy, 1985; Calistri-Yeh, 1991; Eller and Carberry, 1992]. The work reported here takes a different, but complementary, approach: it models how an agent can use what she or he knows about the discourse to recognize whether either participant has misunderstood some previous utterance to repair the misunderstanding. This strategy handles cases that the preventive approaches cannot anticipate. It is also more general, because our system can generate repairs on the basis of the relatively few types of manifestations of misunderstanding, rather than the much broader (and hence more difficult to anticipate) range of sources.

In this paper, we shall describe an abductive account of interpreting speech acts and recognizing misunderstandings (we discuss the generation of repairs of misunderstandings in McRoy and Hirst, 1992). This account is part of a unified theory of speech act production, interpretation, and repair [McRoy, 1993]. According to the theory, speakers use their beliefs about the discourse context and which speech acts are expected to follow from a given speech act in order to select one that accomplishes their goals and then to produce an utterance that performs the chosen speech act. Interpretation and repair attempt to retrace this selection process abductively—when a hearer attempts to interpret an observed utterance, he tries to identify the goals, expectations, or misunderstandings that might have led the to produce it. Previous plan-based approaches [Allen, 1979; Allen, 1983; Litman, 1985; Carberry, 1985] have had difficulty constraining this inference—from only a germ of content, potentially a tremendous number of goals could be inferred. A key assumption of our approach, which follows from insights provided by Conversation Analysis [Garfinkel, 1967; Schegloff and Sacks, 1973], is that participants can rely primarily on expectations derived from social conventions about language use. These expectations enable participants to determine whether the conversation is proceeding smoothly: if nothing unusual is detected, then understanding is presumed to occur. Conversely, when a hearer finds

that a speaker's utterance is inconsistent with his expectations, he may change his interpretation of an earlier turn and generate a repair [Fox, 1987; Suchman, 1987]. Our approach differs from standard CA accounts in that it treats Gricean intentions [Grice, 1957] as part of these conventions and uses them to constrain an agent's expectations; the work thus represents a synthesis of intentional and structural accounts.

Recognizing misunderstanding is like abduction because hearers must explain why, given their knowledge of how differences in understanding are manifested, a speaker might have said what she did. Attributions of misunderstanding are assumptions that might be abduced in constructing such an explanation. Recognizing misunderstanding also resembles a diagnosis in which utterances play the role of "symptoms" and misunderstandings are "faults". Previous work on diagnosis has shown abduction to be a useful characterization [Ahuja and Reggia, 1986; Poole, 1986].

An alternative approach to diagnosing discourse misunderstandings is to reason deductively from a speaker's utterances to his or her goals on the basis of (default) prior beliefs and then rely on belief revision to retract inconsistent interpretations [Cawsey, 1991]; however, this approach has a number of disadvantages. First, any set of rules of this form will be unable to specify all the conditions (such as insincerity) that might also influence the agent's interpretation; a reasoner will need also to assume that there are no "abnormalities" relevant to the participants or the speech event [Poole, 1989]. This approach also ignores the many other possible interpretations that participants might achieve through *negotiation*, independent of their actual beliefs. For example, an agent's response to a yes–no question might treat it as a question, a request, a warning, a test, an insult, a challenge, or just a vacuous statement intended to keep the conversation going. If conversational participants can negotiate such ambiguities, then utterances are at most a *reason* for attributing a certain goal to an agent. That is, they are a *symptom*, not a cause. Any deductive account would thus be counterintuitive, and very likely false as well.

## 2 The abductive framework

We have chosen to develop the proposed account of dialogue using the Prioritized Theorist framework [Poole *et al.*, 1987; Brewka, 1989; van Arragon, 1990]. Theorist typifies what is known as a "proof-based approach" to abduction because it relies on a theorem prover to collect the assumptions that would be needed to prove a given set of observations and to verify their consistency. This framework was selected because of its first-order syntax and its support for both default and abductive reasoning. Within Theorist, we represent linguistic knowledge and the discourse context, and also model how speakers reason about their actions and misunderstandings.

We have used Poole's implementation of Theorist, extended to incorporate preferences among defaults as suggested by Van Arragon [1990]. Poole's Theorist implements a full first-order clausal theorem prover in Prolog. It extends Prolog with a true negation symbol and the contrapositive forms of each clause. Thus, a Theorist clause $\alpha \supset \beta$ is interpreted as $\{\beta \leftarrow \alpha, \neg\alpha \leftarrow \neg\beta\}$. A Prioritized Theorist reasoner can also assume any default $d$ that the programmer has designated as a potential hypothesis, unless it can prove $\neg d$ from some fact or overriding hypothesis.

The reasoning algorithm uses model elimination [Loveland, 1978; Stickel, 1989; Umrigar and Pitchumani, 1985] as its proof strategy. Like Prolog, it is a resolution-based procedure that chains backward from goals to subgoals, using rules of the form $goal \leftarrow subgoal_1 \wedge \cdots \wedge subgoal_n$, to reduce the goals to their subgoals. However, unlike Prolog, it records each subgoal that occurs in the proof tree leading to the current one and checks this list before searching the knowledge base for a relevant clause; this permits it to reason by cases.

## 3 The formal language

The model is based on a sorted first-order language, $\mathcal{L}$, comprising a denumerable set of predicates, variables, constants, and functions, along with the boolean connectives $\vee, \wedge, \neg, \supset$, and $\equiv$, and the predicate $=$. The terms of $\mathcal{L}$ come in six sorts: agents, turns, sequences of turns, actions, descriptions, and suppositions [1]. $\mathcal{L}$ includes an infinite number of variables and function symbols of every sort and arity. We also define a number of special ones: **do, mistake, intend, knowif, knowref, knowsBetterRef, not,** and **and**. Each of of these functions takes an agent as its first argument and an action, supposition, or description for each of its other arguments; each of them returns a supposition. The function symbols that return speech acts each take two agents as their first two argument and an action, supposition, or description for each of their other arguments.

For the abductive model, we define a corresponding language $\mathcal{L}_{\mathbf{Th}}$ in the Prioritized Theorist framework. $\mathcal{L}_{\mathbf{Th}}$ includes all the sorts, terms, functions, and predicates of $\mathcal{L}$; however, $\mathcal{L}_{\mathbf{Th}}$ lacks explicit quantification, distinguishes facts from defaults, and associates with each default a priority value. Variable names are understood to be universally quantified in facts and defaults (but existentially quantified in an explanation). Facts are given by "FACT $w$.", where $w$ is a wff. A default can be given either by "DEFAULT $(p, d)$." or "DEFAULT $(p, d) : w$.",

[1] Suppositions represent the propositions that speakers express in a conversation, independent of the truth values that those propositions might have.

where $p$ is a priority value, $d$ is an atomic symbol with only free variables as arguments, and $w$ is a wff. For example, we can express the default that birds normally fly, as:

DEFAULT $(2, birdsFly(b)) : bird(b) \supset fly(b)$.

If $\mathcal{F}$ is the set of facts and $\Delta^p$ is the set of defaults with priority $p$, then an expression DEFAULT$(p, d) : w$ asserts that $d \in \Delta^p$ and $(d \supset w) \in \mathcal{F}$.

# 4 The architecture of the model

In the architecture that we have formulated, producing an utterance is a default, deductive process of choosing both a speech act that meets an agent's communicative and interactional goals and a utterance that will be interpretable as this act in the current context. Utterance interpretation is the complementary (abductive) process of attributing to the speaker communicative and interactional goals by attributing to him or her a discourse-level form that provides a reasonable explanation for an observed utterance in the current context. Social norms delimit the range of responses that a participant may produce without becoming accountable for additional explanation.[2] The attitudes that speakers express provide additional constraints, because speakers are expected not to contradict themselves. We therefore attribute to each agent:

- A theory $\mathcal{T}$ describing his or her linguistic knowledge, including principles of interaction and facts relating linguistic acts.
- A set $\mathcal{B}$ of prior assumptions about the beliefs and goals expressed by the speakers (including assumptions about misunderstanding).
- A set $\mathcal{M}$ of potential assumptions about misunderstandings and meta-planning[3] decisions that agents can make to select among coherent alternatives.

To interpret an utterance **u**, by speaker **s**, the hearer **h** will attempt to solve:

$$\mathcal{T} \cup \mathcal{B} \cup M \vdash utter(\mathbf{s}, \mathbf{h}, \mathbf{u}, \mathbf{ts})$$

for some set $M \subset \mathcal{M}$, where *ts* refers to the current context.

In addition, acts of interpretation and generation update the set of beliefs and goals assumed to be expressed during the discourse. Our current formalization focuses on the problems of identifying how an utterance relates to a context and whether it has been understood. The update of expressed beliefs is handled in the implementation, but outside the formal language.[4]

## 4.1 Speech acts

For simplicity, we represent utterances as surface-level speech acts in the manner first used by Perrault and Allen [1980]. For example, if speaker **m** asks speaker **r** the question "Do you know who's going to that meeting?" we would represent this as: **s-request(m, r, informif(r, m, knowref(r, w)))**. Following Cohen and Levesque [1985], we limit the surface language to the acts **s-request**, **s-inform**, **s-informref**, and **s-informif**. Discourse-level acts include **inform, informif, informref, askref, askif, request, pretell**[5], **testref, testif** and **warn**, and are represented using a similar notation.

## 4.2 Expressed attitudes

We distinguish the beliefs that speakers act as if they have during a course of a conversation from those they might actually have. Most models of discourse incorporate notions of belief and mutual belief to describe what happens when a speaker talks about a proposition, without distinguishing the expressing of belief from believing (see Cohen et al. 1990). However, real belief involves notions of evidence, trustworthiness, and expertise, not accounted for in these models; it is not automatic. Moreover, the beliefs that speakers as if they have need not match their real ones. For example, a speaker might simplify or ignore certain facts that could interfere with the accomplishment of a primary goal [Gutwin and McCalla, 1992]. Speakers need to keep track of what others say, in addition to whether they believe them, because even insincere attitudes can affect the interpretation and production of utterances. Although speakers normally choose to be consistent in the attitudes they express, they can recant if it appears that doing so will lead (or has led) to conversational breakdown.

Following Thomason [1990], we call the contents of the attitudes that speakers express during a dialogue *suppositions* and the attitude itself simply *active*.[6] Thus, when a speaker performs a particular speech act, she activates the linguistic intentions associated with the act, along with a belief that the act has been done. These attitudes do not depend on the

[2] These norms include guidelines such as "If someone asks you a question, you should answer it" or "If someone offers their opinion and you disagree, you should let them know".

[3] Our notion of "meta-planning" is similar to Litman's [1985] use of meta-plans, but we prefer to treat meta-planning as a pattern of inference that is part of the task specification rather than as an action.

[4] A related concern is how an agent's beliefs might change after an utterance has been understood as an act of a particular type. Although we have nothing new to add here, Perrault [1990] shows how Default Logic might be used to address this problem.

[5] A *pretelling* is a preannouncement that says, in effect, "I'm going to tell you something that will surprise you. You might think you know, but you don't."

[6] Supposition differs from belief in that speakers need not distinguish their own suppositions from those of another [Stalnaker, 1972; Thomason, 1990].

speakers' real beliefs.[7]

The following expressions are used to denote suppositions:

- **do**$(s, a)$ expresses that agent $s$ has performed the action $a$;
- **mistake**$(s, a_1, a_2)$ expresses that agent $s$ has mistaken an act $a_1$ for act $a_2$;
- **intend**$(s, p)$ expresses that agent $s$ intends to achieve a situation described by supposition $p$;
- **knowif**$(s, p)$ expresses that the agent $s$ knows whether the proposition named by supposition $p$ is true;
- **knowref**$(s, d)$ expresses that the agent $s$ knows the referent of description $d$;
- **knowsBetterRef**$(s_1, s_2, d)$ expresses that agent $s_1$ has "expert" knowledge about the referent of description $d$, so that if $s_2$ has a different belief about the referent, then $s_2$ is likely to be wrong;[8] and
- **and**$(p_1, p_2)$ expresses the conjunction of suppositions $p_1$ and $p_2$;
- **not**$(p)$ expresses the negation of supposition $p$.[9]

### 4.3 Linguistic knowledge relations

We represent agents' linguistic knowledge with three relations: *decomp*, a binary relation on utterance forms and speech acts; *lintention*, a binary relation on speech acts and suppositions; *lexpectation*, a three-place relation on speech acts, suppositions, and speech acts. The *decomp* relation specifies the speech acts that each utterance form might accomplish. The *lintention* relation specifies the beliefs and intentions that each speech act conventionally expresses. The *lexpectation* relation specifies, for each speech act, which speech acts an agent believing the given condition can expect to follow.

### 4.4 Beliefs and goals

We assume that an agent's beliefs and goals are given explicitly by statements of the form *believe*$(S, P)$ and *hasGoal*$(S, P, TS)$, respectively, where $S$ is an agent, $P$ is a supposition and $TS$ is a turn sequence.

### 4.5 Activation

To represent the dialogue as a whole, including repairs, we introduce the notion of a *turn sequence* and the *activation* of a supposition with respect to a sequence. A turn sequence represents the interpretations of the discourse that a speaker has considered.

Turn sequences are characterized by the following three relations:

- *turnOf*$(ts, t)$ holds if and only if $t$ is a turn in the sequence $ts$;
- *succ*$(t_j, t_i, ts)$ holds if and only if *turnOf*$(ts, t_i)$, *turnOf*$(ts, t_j)$, $t_j$ follows $t_i$ in $ts$, and there is no $t_k$ such that *turnOf*$(ts, t_k)$, *succ*$(t_k, t_i, ts)$, and *succ*$(t_j, t_k, ts)$;
- *focus*$(ts, t)$ holds if $t$ is a distinguished turn upon which the sequence is focused; normally this is the last turn of $ts$.

We also define a successor relation on turn sequences. A turn sequence $TS2$ is a *successor* to turn sequence $TS1$ if $TS2$ is identical to $TS1$ except that $TS2$ has an additional turn $t$ that is not a turn of $TS1$ and that is the successor to the focused turn of $TS1$.

The set of prior assumptions about the beliefs and goals expressed by the participants in a dialogue is represented as the activation of suppositions. For example, an agent **nan** performing an **informref(nan, bob, theTime)** expresses the supposition **do(nan, informref(nan, bob, theTime))** and the Gricean intention,

**and(knowref(nan, theTime),**
**intend(nan, knowref(bob, theTime)))**

given by the *lintention* relation. We assume that an agent will maintain a record of both participants' suppositions, indexed by the turns in which they were expressed. It is represented as a set of statements of the form *expressed*$(P, T)$ or *expressedNot*$(P, T)$ where $P$ is a simple supposition and $T$ is a turn.

Beliefs and intentions that participants express during a turn of a sequence $ts_1$ become and remain active in all sequences that are successors to $ts_1$, unless they are explicitly refuted.

DEFINITION 1: If, according to the interpretation of the conversation represented by turn sequence $TS$ with focused turn $T$, the supposition $P$ was expressed during turn $T$, we say that $P$ becomes *active* with respect to that interpretation and the predicate *active*$(P, TS)$ is derivable:

FACT $\quad expressed(p, t) \wedge focus(ts, t) \supset active(p, ts).$

FACT $\quad expressedNot(p, t) \wedge focus(ts, t) \supset active(\mathbf{not}(p), ts).$

FACT $\quad \neg(active(p, ts) \wedge active(\mathbf{not}(p), ts)).$

If formula $P$ is active within a sequence $TS$, it will remain active until **not**$(P)$ is expressed:

[7] It is essential that these suppositions name propositions independent of their truth values, so that we may represent agents *talking* about knowing and intending without fully analyzing these concepts.

[8] This specialization is needed to capture the pragmatic force of *pretelling*.

[9] The function **not** is distinct from boolean connective $\neg$. It is used to capture the supposition expressed by an agent who says something negative, *e.g.*, "I do not want to go."

FACT $expressed(p,t) \wedge focus(ts,t) \supset \neg activationPersists(\mathbf{not}(p),t)$.

FACT $expressedNot(p,t) \wedge focus(ts,t) \supset \neg activationPersists(p,t)$.

DEFAULT $(1, activationPersists(p,t)) :$
$active(p,ts_i)$
$\wedge\ successorTS(ts_{now}, ts_i)$
$\wedge\ focus(ts_{now}, t)$
$\supset active(p, ts_{now})$.

### 4.6 Expectation

The following definition captures the notion of "expectation".

DEFINITION 2: A discourse-level action $R$ is *expected* by speaker $S$ in turn sequence $TS$ when:

- An action of type $A$ has occurred;
- There is a planning rule corresponding to an adjacency pair $A$–$R$ with condition $C$;
- $S$ believes that $C$;
- The linguistic intentions expressed by $R$ are consistent with $TS$; and
- $R$ has not occurred yet in $TS$.

DEFAULT $(2, expectedReply(p_{do}, p, \mathrm{do}(s_1, a_2), ts))$:
$active(p_{do}, ts)$
$\wedge\ lexpectation(p_{do}, p, \mathrm{do}(s_1, a_2))$
$\wedge\ believe(s_1, p)$
$\wedge\ lintentionsOk(s_1, a_2, ts)$
$\supset expected(s_1, a_2, ts)$.

FACT $active(p_{do}, ts) \supset \neg expectedReply(p_{do}, p, p_{reply}, ts)$.

The predicate *expectedReply* is a default. Although activation might depend on default persistence, activation always takes precedence over expectation because it has a higher priority (on the assumption that memory for suppositions is stronger than expectation).

The predicate $lintentionsOk(S, A, TS)$ is true if speaker $S$ expresses the linguistic intentions of the act $A$ in turn sequence $TS$, and these intentions are consistent with $TS$.

We also introduce a subjunctive form of expectation, which depends only on a speaker's real beliefs:

FACT $lexpectation(\mathrm{do}(s_1, a_1), p, \mathrm{do}(s_2, a_2))$
$\wedge\ believe(s_1, p)$
$\supset wouldEx(s_1, a_1, a_2)$.

### 4.7 Recognizing misunderstandings

When a dialogue proceeds normally, a speaker's utterance can be explained by abducing that a discourse action has been planned using one of a known range of discourse strategies: *plan adoption, acceptance, challenge, repair,* or *closing.* (Figure 1 includes some examples in Theorist.) In cases of apparent misunderstanding, the same explanation process suggests a misunderstanding, rather than a planned act, as the reason for the utterance. To handle these cases, the model needs a theory of the symptoms of a failure to understand [Poole, 1989]. For example, a speaker *S2* might explain an otherwise unexpected response by a speaker *S1* by hypothesizing that *S2* has mistaken some speech act by *S1* for another with a similar decomposition or *S2* might hypothesize that *S1* has misunderstood (see Figure 2). We shall now consider some applications.

## 5 Some applications

This first example (from [Schegloff, 1992]) illustrates both normal interpretation and the recognition of an agent's own misunderstanding:

| | | |
|---|---|---|
| **T1** | **Mother:** | Do you know who's going to that meeting? |
| **T2** | **Russ:** | Who? |
| **T3** | **Mother:** | I don't know. |
| **T4** | **Russ:** | Oh. Probably Mrs. McOwen and probably Mrs. Cadry and some of the teachers. |

The surface-level representation of this conversation is given as the following:

**T1** **m: s-request(m, r, informif(r, m,knowref(r, w)))**

**T2** **r: s-request(r, m, informref(m, r, w))**

**T3** **m: s-inform(m, r, not(knowref(m, w)))**

**T4** **r: s-informref(r, m, w)**

### 5.1 Russ's interpretation of T1 in the meeting example

¿From Russ's perspective, T1 can be explained as a pretelling, an attempt by Mother to get him to ask her who is going. Russ's rules about the relationship between surface forms and speech acts (*decomp*) include that:

FACT $decomp(\mathbf{s\text{-}request}(s_1, s_2, \mathbf{informif}(s_2, s_1, \mathbf{knowref}(s_2, p))), \mathbf{pretell}(s_1, s_2, p))$.

FACT $decomp(\mathbf{s\text{-}request}(s_1, s_2, \mathbf{informif}(s_2, s_1, \mathbf{knowref}(s_2, p))), \mathbf{askref}(s_1, s_2, p))$.

FACT $decomp(\mathbf{s\text{-}request}(s_1, s_2, \mathbf{informif}(s_2, s_1, \mathbf{knowref}(s_2, p))), \mathbf{askif}(s_1, s_2, \mathbf{knowref}(s_2, p)))$.

Russ has linguistic expectation rules for the adjacency pairs *pretell–askref, askref–informref,* and *askif–informif* (as well as for pairs of other types). Russ also has believes that he knows who's going to the meeting, that he knows he knows this, and that Mother's knowledge about the meeting is likely to be

**Utterance Explanation**

FACT $decomp(u, a_1)$
$\wedge\ try(s_1, s_2, a_1, ts)$
$\supset\ utter(s_1, s_2, u, ts)$.

**Planned Actions**

DEFAULT $(2, intendact(s_1, s_2, a_1, ts))$ :
$shouldTry(s_1, s_2, a_1, ts)$
$\supset\ try(s_1, s_2, a_1, ts)$.

**Plan Adoption**

DEFAULT $(3, adopt(s_1, s_2, a_1, a_2, ts))$:
$hasGoal(s_1, do(s_2, a_2), ts)$
$\wedge\ wouldEx(s_1, \mathbf{do}(s_1, a_1), \mathbf{do}(s_2, a_2))$
$\wedge\ lintentionsOk(s_1, a_1, ts)$
$\supset\ shouldTry(s_1, s_2, a_1, ts)$.

"If agent *S1* intends that agent *S2* perform the action *A2* and *A2* is the expected reply to the action *A1*, and it would be coherent for *S1* to perform *A1*, then *S1* should do so."

**Acceptance**

DEFAULT $(2, accept(s_1, a, ts))$:
$expected(s_1, a, ts)$
$\supset\ shouldTry(s_1, s_2, a, ts)$.

"If agent *S1* believes that act *A* is the expected next action, then *S1* should perform *A*."

Figure 1: Theorist rules for producing and interpreting utterances

**Failure to understand**

DEFAULT $(3, selfMis(s_1, s_2, p, a_2, ts))$ :
$active(\mathbf{do}(s_1, a_M), ts)$
$\wedge\ ambiguous(a_M, a_I)$
$\wedge\ lintention(a_2, p_{li})$
$\wedge\ lintention(a_M, p_{li2})$
$\wedge\ inconsistentLI(p_{li}, p_{li2})$
$\wedge\ p = \mathbf{mistake}(s_2, a_I, a_M))$
$\supset\ try(s_1, s_2, a_2, ts)$.

"Speaker *S* might be attempting action *A* in discourse *TS* if: *S* was thought to have performed action $A_M$; but, the linguistic intentions of $A_M$ are inconsistent with those of *A*; acts $A_I$ and $A_M$ have a similar surface form (and hence could be mistaken); and, *H* may have made this mistake."

**Failure to be understood**

DEFAULT $(3, otherMis(s_1, s_2, p, a_2, ts))$ :
$active(\mathbf{do}(s_2, a_I), ts)$
$\wedge\ ambiguous(a_I, a_M)$
$\wedge\ wouldEx(s_1, \mathbf{do}(s_2, a_M), \mathbf{do}(s_1, a_2))$
$\wedge\ p = \mathbf{mistake}(s_1, a_I, a_M))$
$\supset\ try(s_1, s_2, a_2, ts)$.

"Speaker *S* might be attempting action *A* in discourse *TS* if: speaker *H* was thought to have performed action $A_I$; but, acts $A_I$ and $A_M$ have a similar surface form; if *H* had performed $A_M$, *A* would be expected; *S* may express the linguistic intentions of *A*; and, *S* may have made the mistake."

Figure 2: Rules for diagnosing *mis*understanding

better than his own. We assume that he can make default assumptions about what Mother believes and wants:

FACT *believe*(**r, knowref(r, w)**).
FACT *believe*(**r, knowif(r,knowref(r,w))**).
FACT *believe*(**r, knowsBetterRef(m,r,w)**).
DEFAULT $(1, credulousB(p))$ : $believe(\mathbf{m}, p)$.
DEFAULT $(1, credulousH(p, ts))$ : $hasGoal(\mathbf{m}, p, ts)$.

Russ's interpretation of T1 as a pretelling is possible using the meta-plan for plan adoption and the rule for planned action.

1. The proposition
   $hasGoal(\mathbf{m}, \mathbf{do(r, askref(r, m, w))}, \mathbf{ts(0)})$
   may be explained by abducing
   $credulousH$**(do(r,askref(r, m, w)),ts(0))**.
2. An **askref** by Russ would be the expected reply to a **pretell** by Mother:

   $wouldEx$**(m,do(m,pretell(m, r, w)),**
   **do(r,askref(r, m, w)))**

   It would be expected by Mother because:
   - The *lexpectation* relation suggests that she might try to **pretell** in order to get him to produce an **askref**:
     *lexpectation*(**do(m,pretell(m,r,w)),**
     **knowsBetterRef(m,r,w),**
     **do(r,askref(r,m,w)))**
   - Russ may abduce
     $credulousB$**(knowsBetterRef(m, r, w))**
     to explain
     *believe*(**m,knowsBetterRef(m, r, w)**).
3. The discourse context is empty at this point, so the linguistic intentions of pretelling satisfy *lintentionsOk*.

4. Lastly, Russ may assume[10]

   ***adopt*(m, r, pretell(m, r, w), askref(r, m, w), ts(0))**

Thus, the conditions of the plan-adoption meta-rule are satisfied, and Russ can explain ***shouldTry*(m, r, pretell(m, r, w), ts(0))**. This enables him to explain
***try*(m, r, pretell(m, r, w), ts(0))**
as a planned action. Once Russ explains the pretelling, his *decomp* relation and utterance explanation rule allow him to explain the utterance.

### 5.2 Russ's detection of his own misunderstanding in the meeting example

¿From Russ's perspective, the inform-not-knowref that Mother performs in T3 signals a misunderstanding. Assuming T1 is a pretelling, just prior to T3, Russ's model of the discourse corresponds to the following:

***expressed*(do(m, pretell(m, r, w)), 1)**
***expressed*(knowref(m, w), 1)**
***expressed*(knowsBetterRef(m, r, w), 1)**
***expressed*(intend(m, do(m, informref(m, r, w))), 1)**
***expressed*(intend(m, knowref(r, w)), 1)**
***expressed*(do(r, askref(r, m, w)), 2)**
***expressedNot*(knowref(r, w), 2)**
***expressed*(intend(r, knowref(r, w)), 2)**
***expressed*(intend(r, do(m, informref(m, r, w))), 2)**

T3 does not demonstrate acceptance because **inform(m, r, not(knowref(m, w)))** is not coherent with this interpretation of the discourse. This act is incoherent because **not(knowref(m, w))** is among the linguistic intentions of this **inform**, while according to the model ***active*(knowref(m, w),ts(2))**. Thus, it is not the case that:

***lintentionsOk*(m, inform(m, r, not(knowref(m, w))), ts(2))**

As a result, Russ cannot attribute to Mother any expected act, and must attribute a misunderstanding to himself or to her.

Russ may attribute T3 to a self-misunderstanding using the rule for detecting failure to understand. We sketch the proof below.

1. According to the context,
   ***expressed*(do(m,pretell(m,r,w)),0)**.
   And, Russ may assume that the activation of this supposition persists:
   ***activationPersists*(do(m,pretell(m,r,w)),0)**
   ***activationPersists*(do(m,pretell(m,r,w)),1)**
   Thus,
   ***active*(do(m, pretell(m, r, w)), ts(2))**.
2. The acts **pretell** and **askref** have a surface form that is similar,
   **s-request(m,r,informif(r,m,knowref(r,w)))**
   So,
   ***ambiguous*(pretell(m,r,w), askref(m,r,w))**.
3. The linguistic intentions of the pretelling are:

   **and(knowref(m, w), and(knowsBetterRef(m, r, w), and( intend(m, do(m, informref(m, r, w))), intend(m, knowref(r, w)))))**

   The linguistic intentions of inform-not-knowref are

   **and(not(knowref(m, w)), intend(m, knowif(r,not(knowref(m, w))))).**

   But these intentions are inconsistent.
4. Russ may assume

   **selfMis(m,r, mistake(r,askref(m, r, w), pretell(m, r, w)), inform(m, r, not(knowref(m, w))), ts(2)).**

Once Russ explains the inform-not-knowref, his *decomp* relation and utterance explanation rule allow him to explain the utterance.

### 5.3 A case of other-misunderstanding: Speaker A finds that speaker B has misunderstood

We now consider a new example (from McLaughlin [1984]), in which a participant A recognizes that a another participant, B, has mistaken a request in T1 for a test:

**T1** **A:** When is the dinner for Alfred?

**T2** **B:** Is it at seven-thirty?

**T3** **A:** No, *I'm* asking *you.*

**T4** **B:** Oh. I don't know.

The surface-level representation of this conversation is given as the following:

**T1** **a: s-request(a, b, informref(b, a, d))**

**T2** **b: s-request(b, a, informif(a, b, p))**

**T3** **a: s-inform(a, b, intend(a, do(a, askref(a, b, d))))**

**T4** **b: s-inform(b, a, not(knowref(b, d)))**

[10] The only constraint on adopting a plan, is that the result not yet be achieved:

FACT $active(\text{do}(s, a_2), ts)$
$\supset \neg adopt(s_1, s_2, a_1, a_2, ts)$.

A has linguistic expectation rules for the adjacency pairs *pretell–askref*, *askref–informref*, *askif–informif*, and *testref–askif*. A also believes that she does not know the time of the dinner, that B does know the time of the dinner.[11] We assume that A can make default assumptions about what B believes and wants:

FACT *believe*(a, **not(knowref(a,d))**).
FACT *believe*(a, **knowref(b,d)**).
FACT *hasGoal*(a,**do(b,informref(b,a,d))**,**ts(0)**).
DEFAULT $(1, credulousB(p)) : believe(\mathrm{b}, p)$.
DEFAULT $(1, credulousH(p, ts)) : hasGoal(\mathrm{b}, p, ts)$.

¿From A's perspective, after generating T1, her model of the discourse is the following:

*expressed*(**do(a, askref(a, b, d))**, **1**)
*expressedNot*(**knowref(a, d)**, **1**)
*expressed*(**intend(a, knowref(a, d))**, **1**)
*expressed*(**intend(a,**
**do(b, informref(b, a, d)))**, **1**)

According to the *decomp* relation, T2 might be interpretable as **askif(b, a, p)**. However, T2 does not demonstrate acceptance, because there is no *askref–askif* adjacency-pair from which to derive an expectation. T2 is not a plan adoption because A does not believe that B believes that A knows whether the dinner is at seven-thirty. However, there is evidence for misunderstanding, because both information-seeking questions and tests can be formulated as surface requests. Also, T2 is interpretable as a guess and request for confirmation (represented as **askif**), which would be expected after a test. We sketch the proof below.

1. According to the context:
   *expressed*(**do(a, askref(a, b, d))**, **0**).
   A may assume that the activation of this supposition persists:
   *activationPersists*(**do(a, askref(a, b, d))**, **0**).
   Thus, *active*(**do(a,askref(a,b,d))**,**ts(1)**).
2. The acts **askref** and **testref** have a surface form that is similar, namely
   **s-request(a,b,informref(b,a,knowref(b,d)))**.
   So,
   *ambiguous*(**askref(a,b,d)**, **testref(a,b,d)**).
3. An **askif** by B would be the expected reply to a **testref** by A:

   *wouldEx*(**b,do(a,testref(a, b, d)),**
   **do(b,askif(b, a, p))**)

   From A's perspective, it would be expected by B because:

   - The *lexpectation* relation suggests that A might try to produce a **testref** in order to get him to produce an **askif**:

     *lexpectation*(**do(a,testref(a,b,d))**,
     **and(knowref(b,d),**
     **and(knowif(b,p),**
     **and(pred(p,X),**
     **pred(d,X))))**,
     **do(b,askif(b,a,p))**)

     The condition of this rule requires that B believe he knows the referent of description **d** and that **p** asserts that the described property holds of the referent that he knows. For example, if we represent "B knows when the dinner is" as the description
     **knowref(b, the(X, time(dinner, X)))**,
     then the condition requires that
     **knowif(b, time(dinner, q))** for some **q**. This is a gross simplification, but the best that the notation allows.
   - A may assume that B believes the condition of this *lexpectation* by default.

[11] A must believe that B knows when the dinner is for her to have adopted a plan in T1 to produce an **askref** get B to perform the desired **informref**.

## 6 Conclusion

The primary contribution of this work is that it treats misunderstanding and repair as intrinsic to conversants' core language abilities, accounting for them with the same processing mechanisms that underlie normal speech. In particular, it formulates both interpretation and the detection of misunderstandings as explanation problems and models them as abduction.

We have implemented our model in Prolog and the Theorist framework for abduction with Prioritized defaults. Program executions on a Sun-4 for four-turn dialogues take 2 cpu seconds per turn on average.

Directions for future work include extending the model to handle more than one communicative act per turn, misunderstood reference [Heeman and Hirst, 1992], and integrating the account with sentence processing and domain planning.

## Acknowledgements

This work was supported by the University of Toronto and the Natural Sciences and Engineering Research Council of Canada. We thank Ray Reiter for his suggestions regarding abduction; James Allen for his advice; Paul van Arragon and Randy Goebel for their help on using Theorist; Hector Levesque, Mike Gruninger, Sheila McIlraith, Javier Pinto, and Steven Shapiro for their comments on many of the formal aspects of this work; Phil Edmonds, Stephen Green, Diane Horton, Linda Peto, and the other members of the natural language group for their comments; and Suzanne Stevenson for her comments on earlier drafts of this paper.

## References


[Ahuja and Reggia, 1986] Sanjiev B. Ahuja and James A. Reggia. The parsimonious covering model for inexact abductive reasoning in diagnostic systems. In *Recent Developments in the Theory and Applications of Fuzzy Sets. Proceedings of NAFIPS '86 - 1986 Conference of the North American Fuzzy Information Processing Society*, pages 1–20, 1986.

[Allen, 1979] James F. Allen. *A Plan-Based Approach to Speech Act Recognition.* PhD thesis, Department of Computer Science, University of Toronto, Toronto, Canada, 1979. Published as University of Toronto, Department of Computer Science Technical Report No. 131.

[Allen, 1983] James F. Allen. Recognizing intentions from natural language utterances. In Michael Brady, Robert C. Berwick, and James F. Allen, editors, *Computational Models of Discourse*, pages 107–166. The MIT Press, 1983.

[Brewka, 1989] Gerhard Brewka. Preferred subtheories: An extended logical framework for default reasoning. In *Proceedings of the 11th International Joint Conference on Artificial Intelligence*, pages 1043–1048, Detroit, MI, 1989.

[Calistri-Yeh, 1991] Randall J. Calistri-Yeh. Utilizing user models to handle ambiguity and misconceptions in robust plan recognition. *User Modelling and User Adapted Interaction*, 1(4):289–322, 1991.

[Carberry, 1985] Sandra Carberry. *Pragmatics Modeling in Information Systems Interfaces.* PhD thesis, University of Delaware, Newark, Delaware, 1985.

[Cawsey, 1991] Alison J. Cawsey. A belief revision model of repair sequences in dialogue. In Ernesto Costa, editor, *New Directions in Intelligent Tutoring Systems.* Springer Verlag, 1991.

[Cohen and Levesque, 1985] Philip R. Cohen and Hector J. Levesque. Speech acts and rationality. In *23th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, pages 49–60, 1985.

[Cohen *et al.*, 1990] Philip R. Cohen, Jerry Morgan, and Martha Pollack, editors. *Intentions in Communication.* The MIT Press, 1990.

[Eller and Carberry, 1992] Rhonda Eller and Sandra Carberry. A meta-rule approach to flexible plan recognition in dialogue. *User Modelling and User Adapted Interaction*, 2(1–2):27–53, 1992.

[Fox, 1987] Barbara Fox. Interactional reconstruction in real-time language processing. *Cognitive Science*, 11:365–387, 1987.

[Garfinkel, 1967] Harold Garfinkel. *Studies in Ethnomethodology.* Prentice Hall, Englewood Cliffs, NJ, 1967. (Reprinted: Cambridge, England: Polity Press, in association with Basil Blackwell, 1984.).

[Goodman, 1985] Bradley Goodman. Repairing reference identification failures by relaxation. In *The 23rd Annual Meeting of the Association for Computational Linguistics: Proceedings of the Conference*, pages 204–217, Chicago, 1985.

[Grice, 1957] H. P. Grice. Meaning. *The Philosophical Review*, 66:377–388, 1957.

[Gutwin and McCalla, 1992] Carl Gutwin and Gordon McCalla. Would I lie to you? Modelling context and pedagogic misrepresentation in tutorial dialogue. In *30th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, pages 152–158, Newark, DE, 1992.

[Heeman and Hirst, 1992] Peter Heeman and Graeme Hirst. Collaborating on referring expressions. Technical Report 435, Department of Computer Science, University of Rochester, 1992.

[Litman, 1985] Diane J. Litman. *Plan Recognition and Discourse Analysis: An Integrated Approach for Understanding Dialogues.* PhD thesis, Department of Computer Science, University of Rochester, Rochester, NY, 1985. Published as University of Rochester Computer Science Technical Report 170.

[Loveland, 1978] D. W. Loveland. *Automated Theorem Proving: A Logical Basis.* North-Holland, Amsterdam, The Netherlands, 1978.

[McCoy, 1985] Kathleen F. McCoy. The role of perspective in responding to property misconceptions. In *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*, volume 2, pages 791–793, 1985.

[McLaughlin, 1984] Margaret L. McLaughlin. *Conversation: How Talk is Organized.* Sage Publications, Beverly Hills, 1984.

[McRoy and Hirst, 1992] Susan W. McRoy and Graeme Hirst. The repair of speech act misunderstandings by abductive inference. 1992. Submitted for publication.

[McRoy, 1993] Susan W. McRoy. *Abductive Interpretation and Reinterpretation of Natural Language Utterances.* PhD thesis, Department of Computer Science, University of Toronto, Toronto, Canada, 1993. In preparation.

[Perrault and Allen, 1980] C. Raymond Perrault and James F. Allen. A plan-based analysis of indirect speech acts. *Computational Linguistics*, 6:167–183, 1980.

[Perrault, 1990] C. Raymond Perrault. An application of default logic to speech act theory. In Philip R. Cohen, Jerry Morgan, and Martha Pollack, editors, *Intentions in Communication*, pages

161–186. The MIT Press, 1990. An earlier version of this paper was published as Technical Report CSLI-87-90 by the Center for the Study of Language and Information.

[Poole *et al.*, 1987] David Poole, Randy Goebel, and Romas Aleliunas. Theorist: A logical reasoning system for defaults and diagnosis. In Nick Cercone and Gordon McCalla, editors, *The Knowledge Frontier: Essays in the Representation of Knowledge*, pages 331–352. Springer-Verlag, New York, 1987. Also published as Research Report CS-86-06, Faculty of Mathematics, University of Waterloo, February, 1986.

[Poole, 1986] David Poole. Default reasoning and diagnosis as theory formation. Technical Report CS-86-08, Department of Computer Science, University of Waterloo, Waterloo, Ontario, 1986.

[Poole, 1989] David Poole. Normality and faults in logic-based diagnosis. In *Proceedings of the 11th International Joint Conference on Artificial Intelligence*, pages 1304–1310, 1989.

[Schegloff and Sacks, 1973] Emanuel A. Schegloff and Harvey Sacks. Opening up closings. *Semiotica*, 7:289–327, 1973.

[Schegloff, 1992] Emanuel A. Schegloff. Repair after next turn: The last structurally provided defense of intersubjectivity in conversation. *American Journal of Sociology*, 97(5):1295–1345, 1992.

[Stalnaker, 1972] Robert C. Stalnaker. Pragmatics. In *Semantics of Natural Language*, pages 380–397. D. Reidel Publishing Company, Dordrecht, 1972.

[Stickel, 1989] M. E. Stickel. A Prolog technology theorem prover. *Journal of Automated Reasoning*, 4:353–360, 1989.

[Suchman, 1987] Lucy A. Suchman. *Plans and Situated Actions.* Cambridge University Press, Cambridge, UK, 1987.

[Thomason, 1990] Richmond H. Thomason. Propagating epistemic coordination through mutual defaults I. In Rohit Parikh, editor, *Proceedings, Third Conference on Theoretical Aspects of Reasoning about Knowledge (TARK 1990)*, pages 29–39, Pacific Grove, CA, 1990.

[Umrigar and Pitchumani, 1985] Zerksis D. Umrigar and Vijay Pitchumani. An experiment in programming with full first-order logic. In *Symposium of Logic Programming*, Boston, MA, 1985. IEEE Computer Society Press.

[van Arragon, 1990] Paul van Arragon. *Nested Default Reasoning for User Modeling.* PhD thesis, Department of Computer Science, University of Waterloo, Waterloo, Ontario, 1990. Published by the department as Research Report CS-90-25.