ELSEVIER

Artificial Intelligence 85 (1996) 181–243

Artificial Intelligence

The effect of resource limits and task complexity on collaborative planning in dialogue

Marilyn A. Walker *

AT&T Research Laboratories, 600 Mountain Ave., Rm. 2D-441, Murray Hill, NJ 07974, USA

Received January 1995; revised October 1995

Abstract

This paper shows how agents' choice in communicative action can be designed to mitigate the effect of their resource limits in the context of particular features of a collaborative planning task. I first motivate a number of hypotheses about effective language behavior based on a statistical analysis of a corpus of natural collaborative planning dialogues. These hypotheses are then tested in a dialogue testbed whose design is motivated by the corpus analysis. Experiments in the testbed examine the interaction between (1) agents' resource limits in attentional capacity and inferential capacity; (2) agents' choice in communication; and (3) features of communicative tasks that affect task difficulty such as inferential complexity, degree of belief coordination required, and tolerance for errors. The results show that good algorithms for communication must be defined relative to the agents' resource limits and the features of the task. Algorithms that are inefficient for inferentially simple, low coordination or fault tolerant tasks are effective when tasks require coordination or complex inferences, or are fault intolerant. The results provide an explanation for the occurrence of utterances in human dialogues that, prima facie, appear inefficient, and provide the basis for the design of effective algorithms for communicative choice for resource limited agents.

1. Introduction

Agents may engage in conversation for a range of reasons, e.g. to acquire information, to establish a contract, to make a plan, or to be social. At each point in a dialogue, agents must make communicative choices about what to say and how and when to say it. This paper focuses on agents' communicative choice in collaborative planning dialogues, dialogues whose purpose is to discuss and agree on a plan for future action,

* E-mail: walker@research.att.com.

SSDI 0004-3702(95)00114-X

and potentially execute that plan. I will argue that agents' choices in communicative action, their algorithms for language behavior, must be determined with respect to two relatively unexplored factors in models of collaborative planning dialogues: (1) agents' resource limits, such as limits in attentional and inferential capacity; and (2) features of collaborative planning tasks that affect task difficulty, such as inferential complexity, the degree of belief coordination required, and tolerance for errors.

A primary dimension of communicative choice is the degree of explicitness. For example, consider a simple task of agent A and agent B trying to agree on a plan for furnishing a two-room house. Imagine that A wants B to believe the proposition realized by (1) and believes that B can infer this from the propositions realized in (2): [1]

(1) If we agree to put the green couch in the study, we will have a matched pair of furniture in the study.

(2) (a) I propose that we put the green couch in the study.
 (b) We intend to put the green chair in the study.
 (c) Two furniture items of the same color in the same room achieve a matched pair.

In naturally occurring dialogues, A may produce utterances realizing the propositions in (3)-(6), or other variations [30, 109, 133, 141].

(3) A: I propose that we put the green couch in the study.

(4) A: We intend to put the green chair in the study. I propose that we put the green couch in the study.

(5) A: Two furniture items of the same color in the same room achieve a matched pair. We intend to put the green chair in the study. I propose that we put the green couch in the study.

(6) A: I propose that we put the green couch in the study. That will get us a matched pair.

The communicative choices in (3)-(6) illustrate a general fact: for any communicative act, the same effect can be achieved with a range of acts at various levels of explicitness. This raises a key issue: On what basis does A choose among the more or less explicit versions of the proposal in (3)-(6)?

The single constraint that has been suggested elsewhere in the literature is the REDUNDANCY CONSTRAINT: A should not say information that B already knows, or that B could infer. The REDUNDANCY CONSTRAINT appears in the form of simple dictums such as "don't tell people what they already know", as Grice's Quantity Maxim "do not make your contribution more informative than is required" [47] and as constraints on planning operators for the generation and recognition of communicative plans [2, 13, 27, 86, 92]. So, if we assume that B knows (2b) and (2c), then the only possibility for what A can say is (3).

[1] These examples are from the domain of Design-World to be discussed in Section 4 and are abstractions from naturally occurring examples in which the propositions realized here are realized in a number of different ways. Here the focus is on the logical relationships between the content of each proposition: (2a) and (2b) are minor premises and (2c) is the major premise for the inference under discussion.

The REDUNDANCY CONSTRAINT is based on the assumption that agent A should leave implicit any information she believes that B already knows or she believes that B could infer, in other words, that agent B can always "fill in what is missing" by a combination of retrieving facts from memory and making inferences. In Section 2, I will show that agents in natural dialogues consistently violate the REDUNDANCY CONSTRAINT. I will argue that this should not be particularly surprising since the REDUNDANCY CONSTRAINT is based on several simplifying assumptions:

(i) UNLIMITED WORKING MEMORY ASSUMPTION: everything an agent knows is always available for reasoning;
(ii) LOGICAL OMNISCIENCE ASSUMPTION: agents are logically omniscient;
(iii) FEWEST UTTERANCES ASSUMPTION: utterance production is the only process that should be minimized;
(iv) NO AUTONOMY ASSUMPTION: assertions and proposals by agent A are accepted by default by agent B.

When agents are autonomous and resource limited, these assumptions do not always hold, and the problem of communicative choice remains.

The plan for the paper is as follows: Section 2 motivates a number of hypotheses about the relationship of communicative choice, resource limits and task features using evidence from natural collaborative planning dialogues. These hypotheses are the basis of a model of collaborative planning presented in Section 3. Then Section 4 describes how the model is implemented in a testbed for collaborative planning dialogues called Design-World, which supports experiments on the interaction of agents' communicative choice, resource limits, and features of the task. At this point, in Section 4.1, I review the steps of the method applied so far, and motivate the use of simulation as a method for testing the hypotheses. Section 5 presents the experimental results and discusses the extent to which the hypotheses were confirmed, and then Section 6 discusses the theoretical implications of these results and the extent to which they can be generalized to other tasks, agent properties, and communication strategies.

2. Communicative choice in dialogue

Naturally occurring collaborative planning dialogues are design, problem solving, diagnostic or advice giving dialogues [14,21,30,37,49,97,104,128,141,143]. In order to generate hypotheses about the relation of communicative choice to agent properties and task features, this section examines communicative choice in naturally occurring collaborative planning dialogues. Most of the examples discussed below are excerpts from a corpus of dialogues from a radio talk show for financial planning advice [98], [2] but I will also draw on data from collaborative design, collaborative construction, and computer support dialogues [28,139,142,143].

Dialogue, in general, is modeled as a process by which conversants add to what is assumed to be already mutually believed or intended. This set of assumed mutual beliefs

[2] The corpus consists of 55 dialogues from 5 hours of live radio broadcast, where each dialogue ranged in length from 23 to 100 turns.

and intentions is called the DISCOURSE MODEL, or the COMMON GROUND [119, 140]. In collaborative planning dialogues, the conversants are attempting to add mutual beliefs about the current state of the world and mutual beliefs and intentions about a plan for future action to the discourse model. It is obvious that the efficacy of the final plan and the efficiency of the planning process must be affected by agents' algorithms for communicative choice.

However previous work has not systematically varied factors that affect communicative choice, such as resource limits and task complexity. Furthermore, most previous work has been based on the REDUNDANCY CONSTRAINT, and apparently, its concomitant simplifying assumptions (but see [78, 147, 148]).

To explore the relation of communicative choice to effective collaborative planning, the analysis of naturally occurring collaborative planning dialogues in this paper focuses on communicative acts that violate the REDUNDANCY CONSTRAINT. These acts are INFORMATIONALLY REDUNDANT UTTERANCES, IRUs, defined as: [3]

Definition of informational redundancy. An utterance u_i is INFORMATIONALLY REDUNDANT in a discourse situation $\mathcal{S}$

(i) if u_i expresses a proposition p_i, and another utterance u_j that entails p_i has already been said in $\mathcal{S}$;

(ii) if u_i expresses a proposition p_i, and another utterance u_j that presupposes or implicates p_i has already been said in $\mathcal{S}$.

A statistical analysis of the financial advice corpus showed that about 12% of the utterances are IRUs. As mentioned in Section 1, this should not be particularly surprising since the definition of IRUs reflects several simplifying assumptions. For example, the definition reflects the LOGICAL OMNISCIENCE ASSUMPTION because it assumes that all the entailments of propositions uttered in a discourse and certain default inferences from propositions uttered in a discourse become part of the discourse model. [4] The definition reflects the NO AUTONOMY ASSUMPTION because it assumes that merely saying an utterance u_i that expresses a proposition p_i is sufficient for adding p_i to the discourse model. The fact that IRUs occur shows that the simplifying assumptions are not valid.

The distributional analysis suggests that there are at least three functional categories of IRUs:

Communicative functions of IRUs.

- *Attitude*: to provide evidence supporting beliefs about mutual understanding and acceptance.

[3] The first part of the definition is a variation on Hirschberg's definition of redundant [59] which is used in her theory of scalar implicature. This view of information is also the basis of information theoretic work such as [6].

[4] Presuppositions and implicatures are two types of default inferences [44, 46, 69, 82, 124]. The corpus analysis tagged defaults separately from entailments but found no evidence for a functional difference (see [133]).

- *Attention*: to manipulate the locus of attention of the discourse participants by making a proposition salient.
- *Consequence*: to augment the evidence supporting beliefs that certain inferences are licensed.

IRUs have ANTECEDENTS in the dialogue which are the utterances that originally realized the content of the IRU either through direct assertion or by an inferential chain; in the definition above u_j is an antecedent for u_i. The three communicative functions of IRUs were identified by correlations with distributional features based in part on relations between the IRU and its antecedent, such as textual distance, discourse structure relations, and logical relations. The distributional analysis also analyzed utterance features such as the intonational realization of the IRU, the form of the IRU, and the relation of the IRU to adjacent utterances.

Below, I will briefly give examples of each type of IRU.[5] For each type I will explain how the four simplifying assumptions of previous dialogue models predict that the utterance is informationally redundant. Then we will consider hypothetical agent and task properties under which IRUs function as hypothesized above.

2.1. Attitude IRUs

Attitude IRUs provide evidence supporting beliefs about mutual understanding and acceptance by demonstrating the speaker's attitude to an assertion or proposal made by another agent in dialogue. An Attitude IRU, said with a falling intonation typical of a declarative utterance, is given in (7.27), where M repeats what H has asserted. in (7.26). M and H have been discussing how M and her husband can handle funds invested in IRAs (individual retirement accounts). In (7), and in the other naturally occurring examples below, the ANTECEDENTS of the IRUs are *italicized* and the IRUs are in CAPS.

(7) (24) H: That is correct. It could be moved around so that each of you have two thousand.
(25) M: I see.
(26) H: *Without penalty.*
(27) M: WITHOUT PENALTY.
(28) H: Right.
(29) M: And the fact that I have a, an account of my own from a couple of years ago, when I was working, doesn't affect this at all.

[5] Each communicative function given above includes a number of subtypes that will not be represented by these examples. In addition, the hypothesis that IRUs are a rehearsal mechanism, i.e. agents repeat propositions as an aid to memory, is tested in every experiment by the model of attention/working memory. The hypothesis that agents say IRUs because they cannot think of anything else to say (the DEAD AIR hypothesis) was considered in [131], but I as yet have found no evidence to support it. For example, other indications of hesitation or planning what to say, such as disfluencies and long pauses, are not associated with IRUs.

Header:	INFORM(speaker, hearer, proposition),
precondition:	KNOW(speaker, proposition),
want-precondition:	speaker want INFORM(speaker, hearer, proposition),
effect:	KNOW(hearer, proposition).

Fig. 1. Definition of the INFORM plan operator in [2].

The IRU in (27) provides direct evidence that M heard exactly what H said [8,21]. According to arguments elaborated below and elsewhere [131,133], M's response indirectly provides evidence that she accepts and therefore believes what H has asserted.

The classification of (7.27) as an IRU follows from the NO-AUTONOMY ASSUMPTION. The NO-AUTONOMY ASSUMPTION is usually characterized as an agent being "cooperative" or "helpful". For example, the motivation for the plan effect of the INFORM planning operator from [2] in Fig. 1 is that the hearer is cooperative. In other words, a cooperative hearer always accepts and therefore believes (or knows) what the speaker has previously asserted. But if the effect of the INFORM act always goes through, then there is no reason for M to choose to produce an Attitude IRU in (7.27), in response to H's inform in (7.26).

In recent work, the plan effect shown in Fig. 1 is treated as a default [51,67,96,105]. Perrault's belief transfer rule handles inferring the default acceptance of assertions (inform acts), while Grosz and Sidner's conversational default rule CDR2 handles inferring the default acceptance of proposals [51,96]. In both cases, the default inference of acceptance of an assertion of P or a proposal to achieve P depends on the cooperativity of the hearer and on whether or not the hearer previously believed or intended to achieve $\neg P$. However, Attitude IRUs are produced in many situations, where there is no reason for the default not to go through. In advice giving dialogues, the hearer is cooperative and the hearer does not previously believe or intend to achieve $\neg P$, yet Attitude IRUs are common when the caller asks the talk show host a question and then repeats or paraphrases his response to the question with an Attitude IRU.

Clark and Schaefer proposed that Attitude IRUs provide positive evidence of understanding [9,20,21]. They allow for understanding to be implicitly conveyed, but say that the amount of explicit positive evidence should be "sufficient for current purposes". However, Clark and Schaefer do not address the question of belief transfer since they do not distinguish between indicating understanding and indicating acceptance. Furthermore, they make no predictions about what features of current purposes require more or less positive evidence, and thus lead an agent to produce an Attitude IRU.

Thus, neither the addition of defaults nor the positive evidence model makes any predictions about when an agent should produce an Attitude IRU, since the inference of acceptance goes through by default without the Attitude IRU.

In order to explain the function of Attitude IRUs, the NO-AUTONOMY ASSUMPTION must be replaced with the assumption that hearers always either explicitly or implicitly accept or reject each utterance act that is intended to change their beliefs or intentions. In Section 3, these observations are incorporated into a model of collaborative planning dialogue. Results from testing hypotheses related to the choice to produce Attitude

IRUs are presented elsewhere [131, 133, 135], and will not be discussed further in this paper.

2.2. Consequence IRUs

Consequence IRUs make inferences explicit. For example, consider (8.17):

(8) (15) H: Oh no. *IRAs were available as long as you are not a participant in an existing pension.*
(16) J. Oh I see. Well I did work, *I do work for a company that has a pension.*
(17) H: Ahh. THEN YOU'RE NOT ELIGIBLE FOR EIGHTY ONE.

In (8), (8.15) realizes a biconditional inference rule, (8.16) instantiates one of the premises of this rule, and (8.17) realizes an inference that follows from (8.15) and (8.16), for the particular tax year of 1981, by the inference rule of modus tollens.

The definition of (8.17) as an IRU follows from the LOGICAL OMNISCIENCE ASSUMPTION. If all entailments of utterances are automatically added to the discourse model then (8.17) should not occur. However it is well known that neither human nor artificial agents are logically omniscient [45, 55, 63, 73, 95]. Agents might not have enough time to make all the relevant inferences even when they know all the relevant inference rules [72], especially since producing and interpreting speech in real time has heavy planning and processing requirements. Thus plausible hypotheses are that:

- HYPOTH-C1: Agents choose to produce Consequence IRUs to demonstrate that they made the inference that is made explicit.
- HYPOTH-C2: Agents choose to produce Consequence IRUs to ensure that the other agent has access to inferrable information.

These hypotheses are motivated by the fact that agents are not logically omniscient. In addition, in the case of hypothesis C2, agents may choose to produce Consequence IRUs to ensure that the other agents have access to inferrable information in a timely manner, even when, in principle, they believe the other agent is capable of making the inference.

However, much of communicative efficiency relies on the fact that agents do make some inferences from what has been said. Thus plausible refinements of hypotheses C1 and C2 are that:

- HYPOTH-C3: The choice to produce a Consequence IRU is directly related to a measure of "how hard" the inference is.
- HYPOTH-C4: The choice to produce a Consequence IRU is directly related to a measure of "how important" the inference is.
- HYPOTH-C5: The choice to produce a Consequence IRU is directly related to the degree to which the task requires agents to be coordinated on the inferences that they have made.

Confirmation of these hypotheses entails that the FEWEST UTTERANCES ASSUMPTION does not hold whenever processing effort is relevant to achieving the conversational goals.

2.3. Attention IRUs

Attention IRUs manipulate the locus of attention of the discourse participants by making a proposition salient. Attention IRUs often realize facts that are inferentially related to the assertion or proposal that the speaker is making. For example, consider (9) said by agent A to agent B while walking to work:

(9) (a) Let's walk along Walnut St.
(b) IT'S SHORTER.

Agent B already knew that the Walnut Street route was shorter, so, by the REDUNDANCY CONSTRAINT, A should have simply said (9a).

The classification of (9b) as an IRU reflects the UNLIMITED WORKING MEMORY assumption. If everything an agent knows is always available for reasoning, then agents should never make communicative choices to include utterances such as (9b). However, it is well known that human agents have limited attention/working memory [5,91,95], and resource-bounded artificial agents with limited time to access memory also have limited attentional capacity.

If we define SALIENT propositions as those that are accessible to a resource limited agent at a particular point in time [102, 103], then a possible hypothesis is that A said (9b) to provide B with a salient reason to accept A's proposal to walk along Walnut St. Similar observations apply to (10):

(10) (a) Clinton has to take a stand on abortion rights for poor women.
(b) HE'S THE PRESIDENT.

Here (10b) is already known to the discourse participants, but saying it makes it salient. In order to account for (10b) we must modify the specific hypothesis above to reflect the difference in utterance type between (9a) and (10a). Utterance (9a) is a PROPOSAL whereas (10a) is an INFORM. In (10), A said (10b) to provide B with a salient reason to accept A's assertion about Clinton's obligations. Utterance (9b) is a WARRANT for adopting A's proposal in (9a), and (10b) is SUPPORT for belief in A's assertion. [6]

- HYPOTH-A1: Agents produce Attention IRUs to support the processes of deliberating beliefs and intentions.

Hypothesis A1 means that the production of Attention IRUs is a surface manifestation of the fact that agents' limited working memory limits the accessibility of beliefs used as the basis of deliberation. The hypothesis that the function of Attention IRUs is to make information salient to support interpretation and reasoning is formulated in the DISCOURSE INFERENCE CONSTRAINT:

- HYPOTH-A2: There is a DISCOURSE INFERENCE CONSTRAINT whose effect is that inferences in dialogue are derived from propositions that are currently discourse salient (in working memory).

[6] The relationship between these utterances has been characterized as the inference of a discourse relation [89], or the inference of the speaker's intention [92]. Moser and Moore and Hobbs have argued that these two views are functionally equivalent [61,93].

The DISCOURSE INFERENCE CONSTRAINT is quite general since the inferences that A intends B to make may be any inferences related to the dialogue such as logical deductions, commonsense defaults, inferring part of A's plan, or inferring relations such as WARRANT or SUPPORT [92, 134, 141].

A more complex example illustrating the relationship of limited working memory, inferential processing, and agents' communicative choice is dialogue (11). The caller E has been telling H, the talk show host, how all her money is invested and then poses a question in (11.3):

(11) (3) E: . . .and I was wondering—should I continue on with the certificates or,
(4) H: Well it's difficult to tell because we're so far away from any of them. But I would suggest this—if *all of these are 6 month certificates and I presume they are,*
(5) E: *yes,*
(6) H: *then I would like to see you start spreading some of that money around,*
(7) E: uh huh,
(8) H: Now in addition, how old are you?

.

(discussion about retirement investments consisting of 14 utterances)

.

(21) E: uh huh and
(22) (a) H: But as far as the certificates are concerned,
(b) I'D LIKE THEM SPREAD OUT A LITTLE BIT.
(c) THEY'RE ALL 6 MONTH CERTIFICATES.
(23) E: yes,
(24) H: and I don't like putting all my eggs in one basket. . . .

The utterances in (22b) and (22c) realize two propositions established as mutually believed in utterances (4)–(7), thus they are IRUs. Utterance (8) initiates a subdialogue digression about retirement investments. Since the discussion about retirement investments consists of 14 utterances in which the information in (4)–(7) is not discussed, a plausible hypothesis is that, at (22a), H believes that the information expressed in (4)–(7) is no longer salient [136]. However, H expects E to use this information to make two inferences: (1) that having all your investments in 6 month certificates is an INSTANCE OF the negatively evaluated condition of having all your eggs in one basket; and (2) that this is a WARRANT for E to adopt the intention to spread the certificates out a little bit. Here, therefore, we see two types of inferences: a content-based inference, INSTANCE OF, in the first case and a deliberation-based inference, WARRANT, in the second. It appears that H produces IRUs to ensure that these inferences get made and that H is basing his communicative choice on the DISCOURSE INFERENCE CONSTRAINT.

In addition to the naturally occurring examples of Attention IRUs, another source of evidence for the DISCOURSE INFERENCE CONSTRAINT is the distribution of IRUs that make inferences explicit such as the Consequence IRU in dialogue (8), Section 2.2. Fig. 2 contrasts the distribution of Consequence IRUs and paraphrases, which

	Consequence IRUs	Paraphrase IRUs
Antecedents salient	24	43
Antecedents not salient	8	39

Fig. 2. Distribution of Consequence IRUs that make inferences explicit, as compared with paraphrases, according to whether their antecedents are currently salient.

are two different ways in which an IRU can relate logically to the prior discourse.[7] Paraphrases are syntactic or semantic transformations of a single antecedent utterance [65,90]. Inferences are distinguished from paraphrases by requiring the application of a logical inference rule such as modus ponens. A key difference is that inferences have multiple antecedents while paraphrases do not. A priori we would not expect inferences and paraphrases, as two types of entailments, to distribute differently with respect to whether their antecedents are salient.[8] However Fig. 2 shows that INFERENCES are more likely to have salient premises than PARAPHRASES ($\chi^2 = 4.835$, $p < 0.05$, $df = 1$). This distributional fact provides evidence for the DISCOURSE INFERENCE CONSTRAINT because whenever we have evidence that an inference has been made, the premises are likely to be salient.

The data discussed above provide evidence for the DISCOURSE INFERENCE CONSTRAINT, however it is clear that the effect of the constraint is strongly determined by the limits on working memory. In particular, a corollary of the constraint is that inferential complexity can be directly related to the number of premises that must be simultaneously salient for the inference to be made. These hypotheses can be summarized as follows:

- HYPOTH-A3: The choice to produce an Attention IRU is related to the degree of inferential complexity of a task as measured by the number of premises required to make task-related inferences.
- HYPOTH-A4: The choice to produce an Attention IRU is related to the degree to which an agent is resource limited in attentional capacity.

Finally, it is obvious that various tasks can be characterized in terms of the degree of inferential complexity, and that observations about belief coordination similar to those made about Consequence IRUs also apply to Attention IRUs, giving hypothesis A5.

- HYPOTH-A5: The choice to produce an Attention IRU is related to the degree to which the task requires agents to be coordinated on the inferences that they have made.

In the next sections we will see how we can test these hypotheses.

[7] The other categories are repetitions, making implicatures explicit and making presuppositions explicit.

[8] For the corpus analysis, salient utterances are those within the last two turns. This measure is not perfect but it is replicable.

3. Modeling resource limited collaborative planning dialogues

The naturally occurring examples discussed in the previous section gave rise to a number of hypotheses as to the situations in which communicative choices to include IRUs could either improve the efficacy of a collaborative plan or the efficiency of the dialogue by which that plan was constructed. In this section, I will specify the details of a model of collaborative planning that will be used as the basis of the dialogue simulation testbed in which the hypotheses can be tested. In thinking about models of collaborative planning, I have found it useful to consider models in terms of six features:

(i) agent architecture,

(ii) role of resource limits: whether the agents constructing the collaborative plan have limited resources, and thus whether there is an attempt to either maximize or minimize any aspect of resource consumption, and if so which aspects,

(iii) the mutual belief model: whether the function of the dialogue is to establish mutual beliefs, and whether the mutual belief model is binary or allows for defaults in mutual beliefs,

(iv) utterance act types: types of acts available for agents to communicate with other agents and the effects of each act on the cognitive state of the agents and the collaborative planning process,

(v) mixed initiative: whether one agent is the initiator or both agents have equal initiative,[9]

(vi) plan evaluation: how the collaborative plan is evaluated and what factors determine how good the collaborative plan is.

Most accounts of collaborative planning dialogues are not specific about all of these features, although some accounts provide rich models of particular features. For example, Smith and Guinn provide a richer model of mixed initiative than that provided here [53, 117], and there are precise models of how hearers infer the utterance act type or the intention underlying a particular communicative action [1,86,112,127]. However, to my knowledge no previous work has included a specification of the agent architecture, the relationship of the architecture to language behavior, the role of resource limits, and the plan evaluation process. The remainder of this section provides a specification for each of these features.

3.1. Agent architecture, mutual belief and resource limits

Both the agent architecture and the role of resource limits are addressed by adopting an agent architecture based on the IRMA architecture for resource-bounded agents, shown in Fig. 3 [7,99]. The IRMA architecture has not previously been used to model the behavior of agents in dialogue. The basic components of the modified IRMA architecture are:

[9] This is also called "control" [116,143]. Walker and Whittaker and Guinn argued that the distribution of control in natural dialogue is primarily determined by whether information relevant to the task is distributed between the agents or primarily known by one agent [52,139].

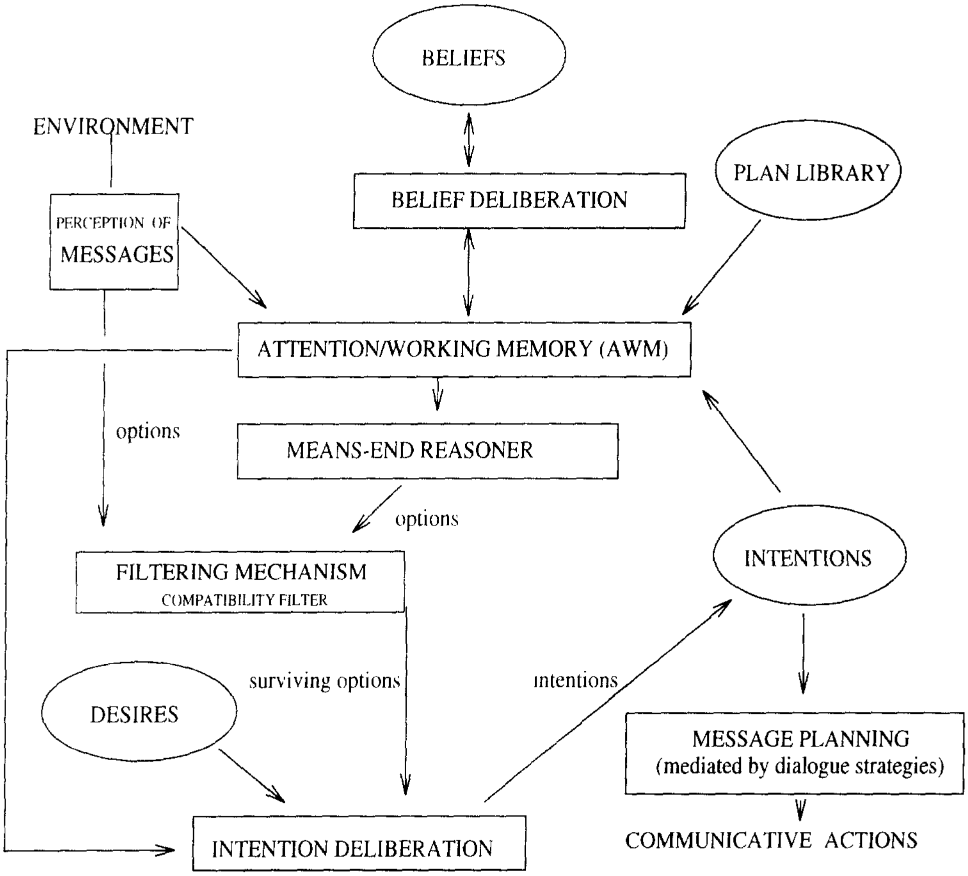

Fig. 3. The IRMA agent architecture for resource-bounded agents with limited attention (AWM).

- Beliefs: a database of an agent's beliefs. This includes beliefs that an agent believes to be mutual to some degree.
- Belief deliberation: decides what an agent wants to believe when there is conflicting evidence.
- Intentions: a database of an agent's intentions. This includes intentions that an agent believes to be mutual to some degree.
- Plan library: what an agent knows about plans as recipes to achieve goals.
- Means-end reasoner: reasons about how to fill in existing partial plans, proposing options that serve as subplans for the plans an agent has in mind.
- Filtering mechanism: checks options for compatibility with the agent's existing plans. Options deemed compatible are passed along to the deliberation process. [10]

[10] The filtering mechanism presented in [7] and used in TileWorld is more complex than that presented here because that work explored the issue of when current intentions get over-ridden.

- Desires: agents may have different types of desires but here I assume that their only desire is to maximize utility [34].
- Intention deliberation: decides which of a set of options to pursue (by an evaluation based on desires such as maximizing utility).
- Attention/working memory (AWM): the limited attention module constrains working memory and the retrieval of current beliefs and intentions that are used by the means-end reasoner.

For the purpose of modeling dialogue, the architecture has been extended with a model of mutual belief that allows for different degrees of mutual belief. For the purpose of exploring the effects of resource bounds on attention, this architecture has been extended with a model of limited attention. All of the modules are standard except for the AWM module described in detail below, and the mutual belief module which will be briefly described.

Attention/working memory model

The model of limited AWM is a cognitively based model adapted from [76], which fits many empirical results on human memory and learning [4,31,57,75,118,121,129]. The motivation for using a cognitively based model of AWM is to model the behavior of agents in naturally occurring dialogues and to test a theory of collaborative communication with humans. [11]

The key properties of the model are that (1) limits on AWM are parameters of the model and can be varied to explore different limits [5]; (2) items encountered more recently are more likely to be salient [76]; (3) items encountered more frequently are more likely to be salient [58].

These recency and frequency effects are a key aspect of the AWM model for testing the hypothesized functions of IRUs. Below I will discuss a particular implementation of this model and its role in testing the hypotheses.

AWM is modeled as a three-dimensional space in which propositions acquired from perceiving the world are stored in chronological sequence according to the location of a moving memory pointer. The sequence of memory loci used for storage constitutes a random walk through memory with each loci a short distance from the previous one. If items are encountered multiple times, they are stored multiple times [58]. The fact that the sequence of storage locations is random means that the recency and frequency effects are stochastically determined. This means that when this model is used in simulation, the simulation produces different results each time.

When an agent retrieves items from memory, search starts from the current pointer location and spreads out in a spherical fashion. The resource limited aspect of AWM follows from the fact that search is restricted to a particular search radius defined in Hamming distance. For example, if the current memory pointer loci is (0 0 0), the loci distance 1 away would be (0 1 0) (0 −1 0) (0 0 1) (0 0 −1) (−1 0 0) (1 0 0). The actual locations are calculated modulo the memory size.

[11] Some of the features of the model hold for processors in general, such as the feature that items that have been discussed more recently are more likely to be accessible with little effort, and that incoming information can displace other information from working memory.

The limit on the search radius defines the subset of the belief and intentions database that is SALIENT. In addition, the fact that the pointer moves has the effect that the salient subset is always changing. Effectively, as new facts are added, others are displaced and become no longer salient, so that the SALIENT predicate is dynamic.

The search radius limit defines the AWM parameter that will be varied in the experiments in Section 5 in order to test the effect of different resource limitations. Experiments that Landauer performed showed that, for a task requiring remembering whether a word belonged to a list of words, the model can be parameterized so that an AWM of 7 reproduces the human performance results in [57]. Since no systematic tests have been performed for human performance on the collaborative planning tasks investigated below, the experiments are run at LOW, MID and HIGH AWM settings. Human performance is assumed to fall somewhere in the middle of these ranges.

Using AWM *to implement the discourse inference constraint*

The model of attention/working memory (AWM) provides a means of testing the hypothesized DISCOURSE INFERENCE CONSTRAINT introduced in Section 2.3. Remember that the discourse inference constraint states that inferences in discourse are derived from propositions that are currently in working memory. The AWM model, as shown in Fig. 3, limits the beliefs accessible for means-end reasoning and deliberation to the subset of beliefs that are currently in AWM [55,66,68,95]. These beliefs are defined as being SALIENT.

This model contrasts with the standard view of inference, where if an agent believes P and believes $P \rightarrow Q$, then the agent believes Q. The discourse inference constraint provides a principled way of limiting inference in modeling humans by requiring the premises of P and $P \rightarrow Q$ to be salient. An axiomatization requires the predicate SALIENT and inference rules as follows for each inference rule schema [61,133]:

Inference under the discourse inference constraint

$$\mathrm{Say}(\mathrm{A},\mathrm{B},P) \rightarrow \mathrm{Salient}(\mathrm{B},P),$$

$$\mathrm{Salient}(\mathrm{B},P) \wedge \mathrm{BEL}(\mathrm{B},P) \wedge \mathrm{Salient}(\mathrm{B},P \rightarrow Q) \wedge \mathrm{BEL}(\mathrm{B},P \rightarrow Q) \\ \rightarrow \mathrm{BEL}(\mathrm{B},Q).$$

The first inference rule states that whenever agent A says an utterance to B that realizes proposition P, that P becomes salient for B. The second rule states that whenever a proposition P and an inference rule, $P \rightarrow Q$, are both believed by B and salient for B, then B can use them to infer Q. The model of AWM must be consulted to determine when the SALIENT predicate holds.

Mutual belief model

The model of mutual belief is based on Lewis' shared environment model of mutual belief [6,19,84], extended to support different degrees of mutual belief by tagging beliefs with qualitative endorsements at the time that they are formed and stored in the beliefs database [29,40,42]. Different degrees of mutual belief allow some actions to be left to inference and some inferences to be defaults. This makes it possible to distinguish between the explicit acceptance of a proposal and the acceptance of a proposal inferred in the absence of evidence to the contrary. When agents are not logically omniscient, it is possible to distinguish between mutual beliefs about what has been mutually inferred and information that has been discussed in the dialogue. See [131, 133] for more detail.

3.2. Discourse acts, utterance acts, and mixed initiative

The overall structure of the discourse in collaborative planning dialogues is primarily determined by the task structure [49, 85, 100, 111]. Each subpart of the task consists of a dialogue segment in which agents negotiate what they should do for that part of the task.

As discussed in Section 2.1, we wish to abandon the NO AUTONOMY ASSUMPTION. The model should allow either agent to initiate the dialogue or initiate a subdialogue about a new part of the task [32, 139]. For each agent to be able to do this, knowledge about the task must be distributed between the participants so that each participant has a basis for means-end reasoning and deliberation.

Furthermore, agents should be able to ACCEPT or REJECT one another's proposals. Each plan step contributing to a higher level goal must remain open for negotiation even if both agents are committed to coming up with a collaborative plan for the higher level goal. This contrasts with models in which proposals for each plan substep that the initiator makes must be accepted by the non-initiator, once the non-initiator has agreed to work on a collaborative plan [26, 51].

Discourse acts and mixed initiative

To engage in collaborative planning, agents take turns sending messages, and each turn may consist of one or more DISCOURSE ACTS. Discourse acts are OPENING, CLOSING, PROPOSAL, ACCEPTANCE, REJECTION and CLARIFICATION. These are higher level acts that are composed of primitives called UTTERANCE ACTS, which will be described below.

The schema of discourse actions shown in Fig. 4 controls the sequence of discourse acts and which discourse acts can be combined into a single turn. [12] The discourse act schema is the basis of an algorithm by which agents achieve a COLLABORATIVE PLAN. For each step in the domain plan:

(i) individual agents perform means-end reasoning about options in the domain;
(ii) individual agents deliberate about which options are preferable;
(iii) then one agent initiates a subdialogue consisting minimally of a PROPOSAL to the other agent, based on the options identified in a reasoning cycle, about actions that contribute to the satisfaction of their goals;
(iv) then the proposal may be subject to CLARIFICATION, after which it is either ACCEPTED or REJECTED by the other agent, by calculating whether it maximizes utility.

[12] This schema cannot describe all discourse action transitions in every type of dialogue [80, 81, 110]. One required extension is to allow multiple proposals to be simultaneously open [107, 113].

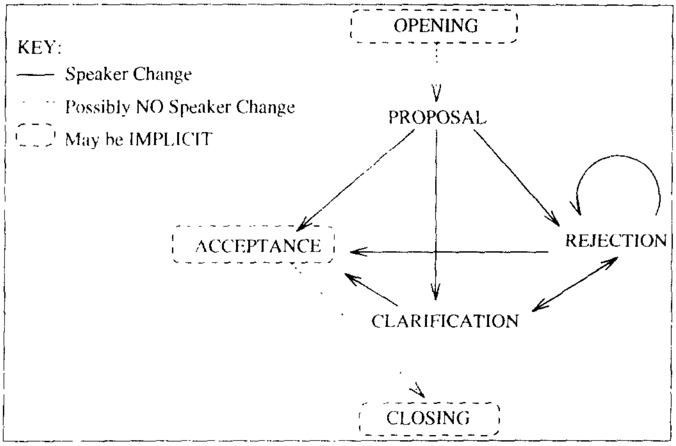

Fig. 4. Finite state model of discourse actions.

This algorithm ties the discourse act schema in Fig. 4 to the IRMA architecture. The requirement that agents must indicate whether they accept or reject each proposal follows from replacing the assumption of cooperativity in earlier work [2, 51] with the COLLABORATIVE PRINCIPLE:

- COLLABORATIVE PRINCIPLE: Conversants must provide evidence of a detected discrepancy in belief as soon as possible.

The COLLABORATIVE PRINCIPLE was proposed in [131], and is an abstraction of the COLLABORATIVE PLANNING PRINCIPLES of Whittaker and Stenton [143] and Walker and Whittaker [139]. The COLLABORATIVE PRINCIPLE means that speakers must monitor the next action by the hearer in order to detect the effects of their utterances. If the hearer continues the dialogue and provides no evidence of a belief discrepancy, the inference of acceptance is licensed as a default [39, 131].

The fact that agents evaluate both assertions and proposals before deciding what to believe or intend follows from the IRMA agent architecture in Fig. 3. As the figure shows, incoming messages about intentions and beliefs are subject to intention or belief deliberation. This provides the basis for abandoning the NO AUTONOMY ASSUMPTION while specifying *why* an agent would accept or reject another agent's proposal (see also [34, 38, 41]). Agents evaluate assertions and proposals from other agents by assessing the support for assertions and the warrants for proposals. Finally, as Fig. 3 shows, this evaluation takes place under constraints of limited working memory, since the beliefs that can serve as supports or warrants must be salient for these processes to use them.

Utterance acts

Fig. 4 shows the discourse act schema that provides the basis for dialogue. Discourse acts are composed of utterance acts, which are the primitive acts that an agent can actually perform. Each discourse act can be performed in different ways by varying the number and type of utterance acts that it consists of. For example, a proposal may or may not include additional information that can convince the hearer, as in example (9).

There are seven utterance act types: OPEN, CLOSE, PROPOSE, ACCEPT, REJECT, ASK and SAY, which are realized via the schemas below:

(Propose ?speaker ?hearer ?option),
(Ask ?speaker ?hearer ?belief),
(Say ?speaker ?hearer ?belief),
(Accept ?speaker ?hearer ?option),
(Reject ?speaker ?hearer ?belief),
(Reject ?speaker ?hearer ?option),
(Open ?speaker ?hearer ?option),
(Close ?speaker ?hearer ?intention).

The content of each utterance act can be an OPTIONS and INTENTIONS representing a domain plan act constructed by means-end reasoning and deliberation as shown in Fig. 3. An OPTION is an act that has not been committed to by both agents. An INTENTION is an act that has been committed to by both agents [7,99]. An option only becomes an intention in the collaborative plan if it is ACCEPTED by both agents, either explicitly or implicitly. The option in a REJECT schema is a counter-proposal and what is rejected is the current proposal.

The content of an utterance act may also be a belief. These beliefs are either those that an agent starts with, beliefs communicated by the other agent, or inferences made by the agent during the conversation. Beliefs in ASK actions have variables that the addressee attempts to instantiate. The belief in a rejection schema is a belief that the speaker believes is a reason to reject the proposal, such as a belief that the preconditions for the option in the proposal do not hold [139].

Examples of these utterance acts in dialogue will be given in Section 4. Below how B processes each of the seven messages that A can send is specified. In the effects specified below, Store means store in AWM, for eventual long-term storage in the beliefs database. The processing involved with each incoming message should be understood with reference to the IRMA agent architecture.

(i) Agent A: (Propose ?speaker ?hearer ?option).
Agent B:
  (a) Filter: check whether ?option is compatible with current beliefs, e.g. that no current beliefs contradict its preconditions.
  (b) Infer and store the preconditions of ?option.
  (c) Means-end reason (ME-reason) about intention the ?option contributes to.
  (d) Deliberate by evaluating the ?option against other options generated by means-end reasoning.
  (e) Indicate results of deliberation by an accept or reject.

(ii) Agent A: (Ask ?speaker ?hearer ?belief).
Agent B: Retrieve beliefs matching ?belief from memory and respond with (say ?speaker ?hearer ?belief) with the variable instantiated for each matching Belief.

(iii) Agent A: (Say ?speaker ?hearer ?belief).
Agent B: Store ?belief.
(iv) Agent A: (Accept ?speaker ?hearer ?option).[13]
Agent B:
(a) Store (intend A B ?option).[14]
(b) Store (act-effects ?option).
(v) Agent A: (Reject ?speaker ?hearer ?option).
Agent B:
(a) Deliberate ?option in comparison with own current proposal that was rejected.
(b) Accept ?option if better than your current proposal.
(c) If rejecting ?option then reject with reason for rejection.
(vi) Agent A: (Reject ?speaker ?hearer ?belief).
Agent B: Store ?belief.
(vii) Agent A: (Open ?speaker ?hearer ?option).
Agent B: Mark the discourse segment that matches ?option as open.
(viii) Agent A: (Close ?speaker ?hearer ?intention).
Agent B: Close the discourse segment for ?intention.

These acts and their effects determine the structure of the dialogue and its effect on the mental state of the conversants.

3.3. *Plan evaluation*

There are three components of the plan evaluation process that have different effects on collaborative planning. Two features are related to the task definition and the third to the model of evaluation applied:

(i) the degree of belief coordination: whether some or all of the *intentions* associated with a plan must be mutually intended and whether any *beliefs* related to the intended acts must also be mutually believed;
(ii) task determinacy and fault tolerance: whether the task has only one solution, or is fault tolerant or more or less satisfiable;
(iii) the model must specify what is to be optimized and whose resource consumption is to be minimized for performance evaluation.

Different theories of collaborative planning reflect different views of the degree of belief coordination required for agents have a collaborative plan. The minimal approach is to not require the agents to establish mutual beliefs at all [35,53]. Rather agents divide up the plan into subcomponents and separately plan each component, without requiring agreement on how the subcomponents are planned. At the next level of belief coordination, it is common to require the intended acts of the collaborative plan to be mutually intended [51,79,125,127]. At the highest level of belief coordination, the

[13] This is a simplification since the form of the acceptance determines the endorsement type on the mutual belief that is added to the beliefs database.

[14] This represents that both agents are committed to the option while the binding of the ?option specifies the agent who will execute the option.

agents must both mutually intend all intentions and mutually believe any beliefs that support the plan such as the WARRANT beliefs that provide reasons for adopting a step of the plan. In addition, it is possible to require that inferences about other goals that the intended actions will satisfy should also be mutually believed.

In this work, the assumption is that the degree of belief coordination required is a feature of the task. The minimal level in the experiments discussed below will be that all intentions must be mutually intended, and the experiments will vary whether WARRANTS, and inferred intentions that are derived from explicitly discussed intentions, must also be mutual.

Task determinacy and fault tolerance have an effect on communicative choice in collaborative planning because they are directly related to how much uncertainty is tolerable in the plan. If a partial plan has some utility, then making a mistake or only constructing a partial plan is not catastrophic. For task determinacy, I assume that a measure of the quality of the final plan can be determined from the utility of each step in the plan, and that partial plans can also be evaluated, so that the task is more or less satisfiable.

With respect to evaluating performance, I assume that the agents in a collaborative planning dialogue are working as a team, and as a team they attempt to optimize the team's performance and minimize the team's consumption of resources. This follows from Clark's assumption that conversants in dialogue attempt to achieve their dialogue purpose with LEAST COLLABORATIVE EFFORT [18,21,22]. This approach contrasts with other approaches in which agents only participate in communication to the degree that it maximizes their own expected utility [35].

A final choice has to do with which processes collaborative effort consists of. A common assumption is that the number of utterances is the primary efficiency measure [15,47]; this is the FEWEST UTTERANCES ASSUMPTION. Since all types of IRUs violate this assumption, in this work collaborative effort is defined with reference to the agent architecture and to all the processes required in collaborative planning, i.e. (1) retrieval processes necessary to access previously stored beliefs in memory; (2) communicative processes related to generating and interpreting utterances; and (3) reasoning processes that operate on beliefs stored in memory and those communicated by other agents. With respect to the IRMA architecture (Fig. 3), retrieval processes are those that access AWM, the plan library and the beliefs and intentions databases, communicative processes are the modules for perception and generation of messages, and inferences are the combined processes of deliberation, means-end reasoning, and filtering. Collaborative effort includes the costs for both agents for all of these processes:

$$\begin{aligned}\text{COLLABORATIVE EFFORT} = &\ (\text{the total cost of communication for both agents})\\ &+(\text{the total cost of inferences for both agents})\\ &+(\text{the total cost of retrievals for both agents}).\end{aligned}$$

Collaborative effort is defined for the whole dialogue and not on a per utterance basis. This definition and the other assumptions support the specification of the plan evaluation process. Given the above definitions, performance is the difference between a measure of the quality of the problem solution and COLLABORATIVE EFFORT.

PERFORMANCE = QUALITY OF SOLUTION − COLLABORATIVE EFFORT.

Since the agents' desires are simply to maximize utility, the quality of the solution is measured by the utility of the resulting plan with respect to the agents' utility functions.

4. Design-World

4.1. Methodological basis of Design-World

Design-World is a testbed for a theory of collaborative communication, which instantiates the model of collaborative planning in dialogue discussed in Section 3. In order to motivate the use of the Design-World testbed in developing and testing a defeasible theory of communication in collaborative planning, this section first describes Design-World as an instance of a general method, and then describes the testbed and its implementation as well as the task and communication parameters. The method can be characterized by the steps below:

(i) Generate hypotheses about the features of a model of collaborative planning dialogues from a statistical analysis of human–human dialogue corpora.
(ii) Produce a functional characterization of the model, specifically including the parameters that could affect task outcome, or claims about the efficacy of the model.
(iii) Implement the model as a testbed so that (some of) these parameters can be controlled, while using independently motivated modules for other aspects of testbed.
(iv) Test the hypotheses and the resulting model against different situations controlled by parameter settings.

The hypotheses that were generated by the statistical analysis of the dialogue corpora were discussed in Section 2. These hypotheses are roughly that under constraints of resource bounds, task inferential complexity, task fault tolerance, and task requirements for belief coordination, communicative choices to include IRUs can reduce collaborative effort or increase the quality of solution of the collaborative plan.

The next step is to produce a functional characterization of the model (a partial formalization). In Section 3, I discussed features of a model of collaborative planning that interact with an agent's autonomy, resource limits and communicative choices. In non-experimental work, the model is the final result of the research. However, this leaves the model and the claims that motivated the model empirically unverified. In formal characterizations, many simplifying assumptions need to be made, and it is not always clear that the results carry over to complex domains where the simplifying assumptions do not hold. While the model presented here is empirically based on statistical analysis of a corpus of naturally occurring dialogues, many of the hypotheses discussed above are related to models of agents' processing. Corpus analysis can only provide weak support for these hypotheses. Thus another source of empirical verification is desirable in order to develop a well-specified and defeasible theory. This is the motivation for the Design-World testbed.

Next, it is necessary to consider the parameters that could affect the outcome or claims about the efficacy of the model and then implement the model as a testbed so that at least some of these parameters can be controlled. The use of independently motivated modules for other aspects of the testbed guarantees that the testbed actually tests something, and also makes it less likely that the testbed is a case of "experimentation in the small" [54].

Parameters that affect the efficacious use of IRUs have already been discussed: these include resource bounds, task inferential complexity, and requirements for belief coordination. The AWM model introduces a parameter for resource bounds, and is implemented as part of the IRMA architecture for resource limited agents, which is independently motivated [7,99]. The AWM model itself is also independently motivated, since it reproduces in simulation many well-known results on human memory and learning [76, 118]. In addition, the utterance acts and their effects are independently motivated by other research on collaborative planning dialogues and by the statistical analysis of dialogue corpora [12, 17, 106, 113, 120, 127, 139, 143].

To introduce parameters related to tasks, the testbed is designed around a simple task where two agents must form a collaborative plan as to how to arrange some furniture in the rooms of a two-room house. The task is based on cooperative design tasks used for experiments on distributed cooperative work for which a corpus of dialogues was available [142]. However, the simple task can be varied along three dimensions: (1) inferential complexity; (2) degree of belief coordination required; (3) tolerance for errors and usefulness of partial solutions.

These three task dimensions represent very different tasks. For example, varying the task by increasing inferential complexity provides information on the performance of agent communication algorithms in simple versus inferentially complex tasks. The dimensions enable us to generalize from the specific task in Design-World to real-world tasks. Section 4.2 will introduce the Standard version of the task, and then Section 4.4 will describe the task variations.

To introduce parameters related to the interaction of communicative choice with task complexity and resource limits, agents are designed so that they vary their communication strategies to include or not include IRUs. Section 4.5 will describe the communicative choice parameters that will be used in the experimental results presented in Section 5.

The experiments reported in Section 5 will examine the interaction of three factors: (1) resource limits; (2) communicative strategies; and (3) task definition. The experiments in the testbed have several functions: (1) they demonstrate that the model can be implemented; (2) they highlight potential flaws in the model; and (3) they provide empirical verification of hypotheses about the function of particular communicative strategies beyond that provided by corpus analysis and researcher's intuitions. This section describes the domain, the implementation of the collaborative planning model in this domain, the communicative strategies and the task variations.

4.2. *Design-World collaborative planning domain*

In Design-World, two artificial parameterizable agents converse in order to agree on the design of the floor plan of a two-room house [133, 142]. The DESIGN-HOUSE plan

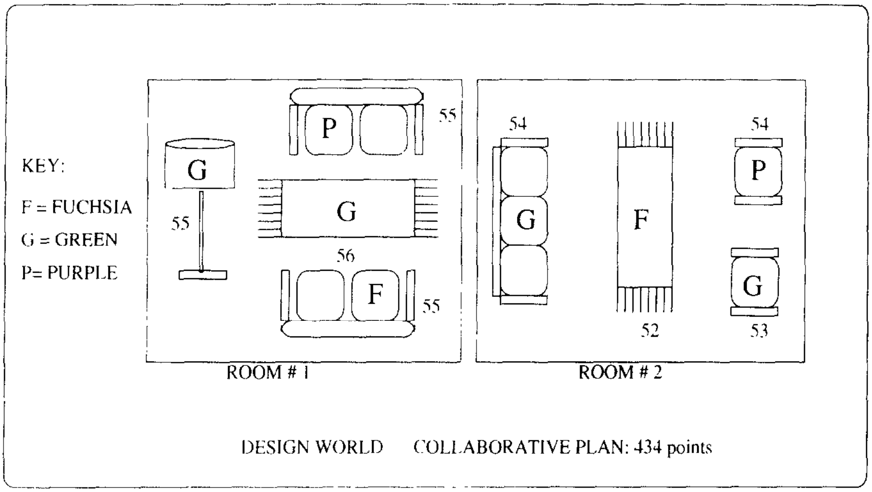

Fig. 5. One final state for Design-World Standard task: represents the collaborative plan achieved by the dialogue, 434 points.

requires the agents to agree on how to DESIGN-ROOM-1 and DESIGN-ROOM-2. At the beginning of the simulation, both agents know the structure of the DESIGN-HOUSE plan and that it requires four furniture items in each room. Each agent has 12 items of furniture that can be used in the plan. A potential final collaborative design plan negotiated via a (simulated) dialogue is shown in Fig. 5.

Each furniture item has a value that is used both as a WARRANT to support deliberation (the utility of an act that is under consideration), and as the basis for calculating the QUALITY OF SOLUTION of the final plan, as discussed in Section 3. The values on the furniture items range from 10 to 56, and both agents' furniture items range over these values. Since beliefs about these values will be used to test hypotheses about the function of IRUs, both agents know the values of all the furniture items at the beginning of the dialogue.

Design-World agents negotiate a collaborative plan following the collaborative planning algorithm discussed in Section 3. Each agent carries out means-end reasoning about the furniture items that can be used in the design. Means-end reasoning generates OPTIONS—these options are the content of PROPOSALS to the other agent to PUT a furniture item into a room. This is illustrated by dialogue (12), which includes the artificial language that the agents communicate with and a gloss automatically generated from that language in *italics* for part of designing room-1.

(12) (1) BILL: *First, put the green rug in the study.*
(Propose agent-bill agent-kim option-10:
put-act (agent-bill green rug room-1))
(2) KIM: *Then, let's put the green lamp in the study.*
(Propose agent-kim agent-bill option-33:
put-act (agent-kim green lamp room-1))

(3) BILL: *Then, let's put the green couch in the study.*
(Propose agent-bill agent-kim option-45:
put-act (agent-bill green couch room-1))
(4) KIM: *No, instead let's put the purple couch in the study.*
(Reject agent-kim agent-bill option-56:
put-act (agent-kim purple couch room-1))
(5) BILL: PUTTING IN THE GREEN COUCH IS WORTH 56.
(Say agent-bill agent-kim bel-265:
score (option-45: put-act (agent-bill green couch room-1) 56))
(6) BILL: *It is better to put the green couch in the study.*
(Reject agent-bill agent-kim option-56:
put-act (agent-bill green couch room-1))

At the beginning of the dialogue, agent-kim has stored in memory the proposition that (score green rug 56). When she receives Bill's proposal as shown in (12.1), she evaluates that proposal in order to decide whether to accept or reject it. As part of evaluating the proposal she will attempt to retrieve the score proposition stored earlier in memory. Thus the propositions about the scores of furniture items are WARRANTS for supporting deliberation.

As discussed in Section 3, the agents retain their autonomy even though the agents both want to agree on a plan for designing the house. Thus, on receiving a proposal, an agent deliberates whether to ACCEPT or REJECT it [34, 135]. Proposals (1) and (2) are inferred to be implicitly ACCEPTED because they are not rejected [139, 143]. This follows from the COLLABORATIVE PRINCIPLE discussed in Section 3. If a proposal is ACCEPTED, either implicitly or explicitly, then the option that was the content of the proposal becomes a mutual intention that contributes to the final design plan [101, 113, 131].

Agents REJECT a proposal if deliberation leads them to believe that they know of a better option, based on evaluating the utility of the competing options they have generated by means-end reasoning. For example, in (12.4) Kim rejects the proposal in (12.3), for pursuing option-45, and proposes option-56 instead. The form of the rejection as a counter-proposal is based on observations about how rejection is communicated in naturally occurring dialogue as codified in the COLLABORATIVE PLANNING PRINCIPLES [139]. When an agent intends to reject another agent's rejection, as in (12.5) and (12.6), the agent includes additional information to support its proposal. In (12.5), agent-bill reminds agent-kim of the value of the green couch, before rejecting agent-kim's proposal.

4.3. Agent architecture implementation in Design-World

The agent architecture used in the Design-World simulation environment is the modified IRMA architecture, shown in Fig. 3 and discussed in Section 3 [7, 99]. The only aspects of the architecture that are specific to Design-World are the plan library, the way that AWM is implemented, and the way that belief deliberation is implemented.

For the experiments below, the total size of AWM is set to 16, but memory is wrap-around, and there is no overwriting. If the path of the memory pointer retraces its steps so that the current memory loci already has something stored in it, the new item is simply added. Thus memory capacity is unbounded.

Since hypothesis A4 relates to the degree to which AWM is limited, we want to be able to compare the performance of agents who are more or less attention limited. Thus, all experiments make comparisons between different communicative strategies over three ranges of AWM settings; the AWM search radius parameter varies from LOW AWM (radius of 3 and 4), to MID AWM (radius of 6 and 7) to HIGH AWM (radius of 11 and 16). LOW AWM agents are severely attention limited agents, wherease almost everything an agent knows is salient for HIGH AWM agents.

The limits on AWM plays a critical role in determining agents' performance. Remember that only *salient* beliefs can be used in means-end reasoning and deliberation, so that if the warrant for a proposal is not salient, the agent cannot properly evaluate a proposal. However, if the agent only knows of one option, the agent can accept the proposal on the assumption that any option is better than doing nothing. Section 5 will show the impact of resource limits on performance. For more details see [76, 135].

The implementation of belief deliberation, for the purpose of Design-World, was tied directly to Landauer's AWM model. In that model, nothing stored in memory is ever deleted or modified. Rather, new beliefs are added which effectively compete with beliefs that are already present. Thus an agent's belief deliberation process depends on collecting a set of related beliefs which may be contradictory, and applying an algorithm to determine what the agent believes (see [39, 43]). The fact that new information about the state of the world supercedes old information is an emergent property of the belief retrieval mechanism: beliefs recently added are more likely to be retrieved. However the stochastic aspect of retrieval means that it is possible for an agent to decide to believe "out-of-date" propositions, and "forget" recent changes in the world. As we will see in Section 5, this means that, in cases where the outdated beliefs were encountered with greater frequency and thus stored in memory repeatedly, agents who can access all of their beliefs are more likely to decide to believe out of date beliefs. [15]

The plan library contains domain plans for Design-house and its subgoals, as well as discourse plans for the discourse acts shown in Fig. 4. The discourse plans will be discussed in detail in Section 4.5.

4.4. Design-World tasks

The Design-World task as a plan is simple since it involves linear planning of subgoals which contribute to higher level goals, as shown in Fig. 6. However, the task is easily modified according to the general task features discussed above so that it is more difficult to perform well. These modifications are applicable to other tasks besides the testbed task, and affect the degree to which different aspects of the task contribute to the performance evaluation.

[15] It is unclear whether this prediction of the belief deliberation algorithm is consistent with human performance. However it is easy to think of examples of humans making the kind of error that this model would predict. For example, I commonly believe (falsely) that I have eggs at home in the refrigerator, even though I used them the previous evening for quiche

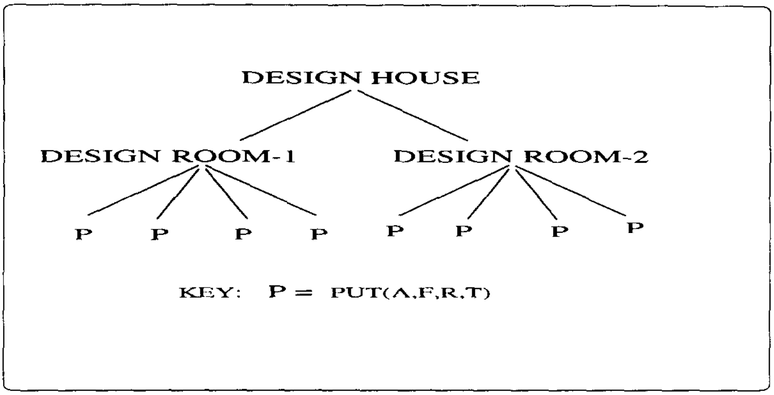

Fig. 6. Standard version of the task, fault tolerant and partial solutions acceptable.

There are four versions of the task that will be used to test the hypotheses introduced in Section 2: Standard, Zero-non-matching-beliefs, Matched-pair, and Zero-invalids. The Standard task is inferentially simple, fault tolerant and requires low levels of belief coordination. The other tasks are more difficult because they increase the degree of belief coordination required, and magnify the effect of mistakes. The Zero-non-matching-beliefs and Matched-pair tasks each explore different aspects of belief coordination and inferential complexity. The Zero-invalids task is fault intolerant.

Standard task

The Standard task provides a baseline and is inferentially simple, fault tolerant and requires low levels of belief coordination. The Standard task is defined so that the QUALITY OF SOLUTION for a particular dialogue consists of the sum of all the furniture pieces for each valid step in the plan. In addition, the task is defined so that partial solutions are possible. Any number of furniture items in a room is a valid plan, rather than requiring that each room *must* have all four furniture items. This choice about task determinacy makes it possible to see the gradient effect on performance of different resource restrictions.

In addition the Standard task is fault tolerant. If agents make a mistake in planning and insert invalid steps in their collaborative plan, the point values for invalid steps in the plan are simply subtracted from the score. Thus in the Standard task agents are not heavily penalized for making mistakes due to inserting steps in plans that are not actually executable.

The Standard task is inferentially simple because the agent's only inferences are those by means-end reasoning to generate options, those by deliberation to evaluate options, and act-effect inferences after committing to an action. Each of these inferences relies on one premise: the premise (has ?agent ?item) supports means-end reasoning, the premise (score ?item ?score) supports deliberation, and the premise (intend A B ?option) supports inferring the effect of ?option. Thus, none of these processes requires multiple premises to be simultaneously salient. However, it is possible to test hypotheses about processing effort in the Standard task by making it easier to access these inferential

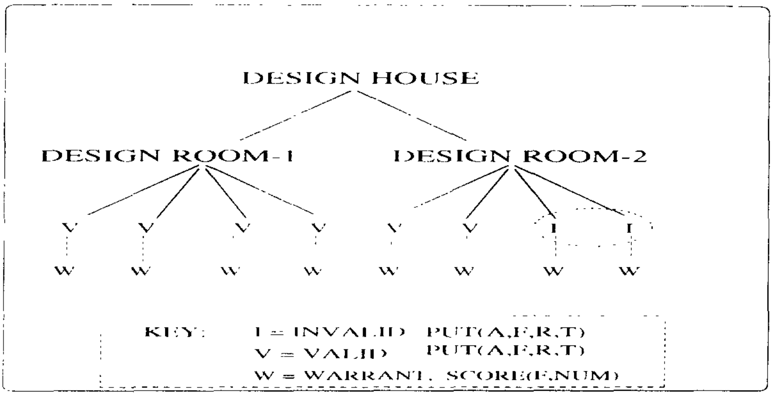

Fig. 7. Tasks can differ as to the level of mutual belief required. Some tasks require that W, a reason for doing P, is mutually believed and others don't.

premises. It is also possible to test the effect of resource limits since these premises must be accessible to perform optimally on the task.

The degree of belief coordination in the Standard task is low because agents are only required to coordinate on the intentions corresponding to put-acts as shown in Fig. 6. These intentions are always explicitly discussed so that coordination is always achieved.

The Zero-non-matching-beliefs task

The Zero-non-matching-beliefs task increases the degree of belief coordination by requiring agents to base their deliberation process on the same beliefs. They must have the same WARRANTS for adopting an intention in order to do well on this task. Fig. 7 shows the structure of beliefs about intentions and warrants for the Design-house goal. In the Zero-non-matching-beliefs task, as shown, the warrants underlying intentions must also be mutually believed. This is not generally required in forming a collaborative plan because agents A and B can mutually believe that they have maximized utility without necessarily agreeing on what that utility is. Furthermore, in the general case, when agents have only one option under consideration, they do not need to evaluate the utility of that one option in order to decide whether to accept or reject it.

The Zero-non-matching-beliefs task provides a basis for testing hypotheses A1 and A5, introduced in Section 2 by increasing the degree of belief coordination required to perform well on the task, where the beliefs are those used in deliberation.

The Zero-non-matching-beliefs task models particular types of real-world tasks since it is not always necessary for agents to agree on the reasons for carrying out a particular action. For example, in the negotiation between the union and the management of a company, any agreement that is reached is agreed to by each party for different reasons. An agreement for a shorter work week is supported by the union because more overtime pay is possible for those who want to work more and is supported by the management because the company's insurance premiums will be lower. However, if two agents agree on a plan, but have different reasons for doing so, they may change their beliefs and

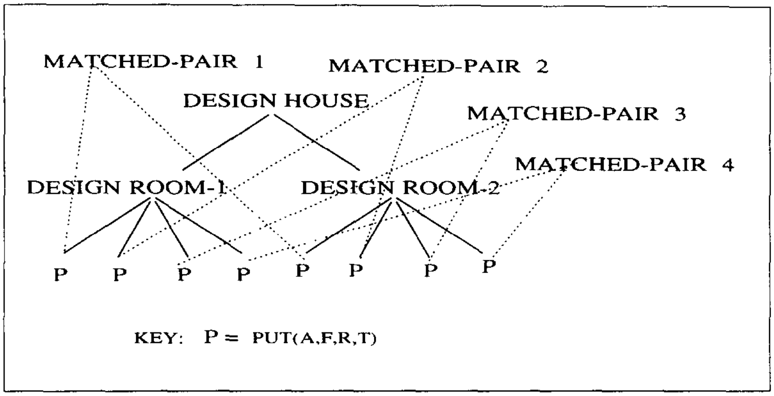

Fig. 8. Making additional inferences: Matched-pair-two-room task. Each PUT intention contributes both to a Design-room goal as well as a Matched-pair goal.

their intentions under different conditions. The most stable, long-term, collaborative plans will be those in which agents agree on both the actions to be performed, as well as the reasons for doing those actions. Under these conditions the agents will be more likely to revise their intentions in a compatible way and intention revision should be simpler. Thus the Zero-non-matching-beliefs task examines one extreme of belief coordination for deliberation.

Matched-pair tasks

Another aspect of belief coordination has to do with coordinating beliefs based on inferences. There are two task definitions that increase inferential complexity by increasing the number of independent premises that must be simultaneously available in working memory. These are: (1) Matched-pair-same-room, and (2) Matched-pair-two-room. Fig. 8 shows the Matched-pair-two-room version of the task. Each intention to PUT a furniture item in a room can potentially contribute to another intention of achieving a matched pair goal. A Matched-pair is two furniture items of the same color. The inference of a Matched-pair is based on the minor premises shown in (13):

(13) (a) (Intend A B (Put ?agent ?item-1 ?room-1))
(b) (Intend A B (Put ?agent ?item-2 ?room-2))
(c) (Equal (Color ?item-1) (Color ?item-2))

Making this inference is more demanding for resource limited agents than the processing needed in the Standard task. In the Standard task, in order to agree on one step of the plan, the agents must access at least one belief about a furniture item they have available. In order to properly evaluate the option represented by that furniture item they must access the WARRANT for that option. In contrast, in both Matched-pair tasks, the WARRANT for both beliefs must be accessed for both furniture items that could contribute to a Matched-pair goal. In addition, the premises in (13) must also be accessed, and furthermore, each version of the Matched-pair task requires one additional premise.

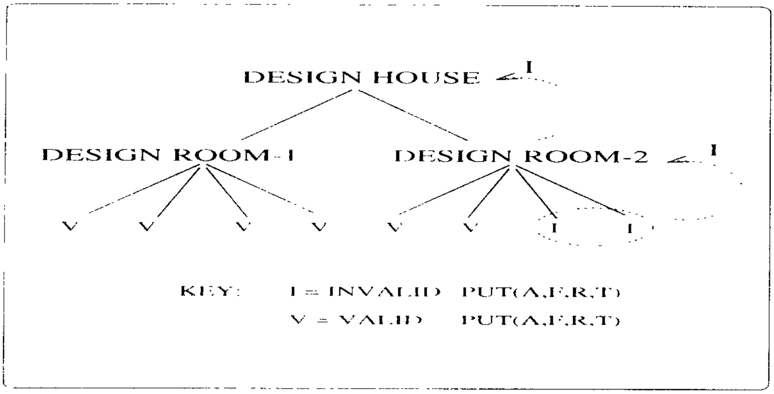

Fig. 9. In Zero-invalids task, invalid steps invalidate the whole plan.

The difference in the two Matched-pair tasks is whether the matches are in the same or different rooms. Matched-pair-same-room requires the additional premise that (Equal ?room-1 ?room-2) while Matched-pair-two-room requires the additional premise that NOT (Equal ?room-1 ?room-2). Because premise (13a) is inferred and stored in memory at the time that a proposal is accepted, and because the agents always complete one room before starting another, the necessary premise shown in (13a) is more likely to be salient in the Matched-pair-same-room task.

As discussed in Sections 1 and 2, we wish to provide a test of hypothesis A2, the discourse inference constraint, and examine how it affects the coordination of inference in collaborative planning. Hypotheses A3 and A5 together imply that the complexity of inference should interact with the agent's ability to stay coordinated on inferences. Since our measure of inferential complexity is the number of independent premises required to draw an inference, both the Matched-pair-same-room task and the Matched-pair-two-room task increase inferential complexity.

Evaluating the QUALITY OF SOLUTION for the Matched-pair tasks reflects the emphasis on coordinating on inferences, since both Matched-pair tasks require that *both* agents make the Matched-pair inferences in order to score points for matched pairs. The task measures how well agents are *coordinated* on the inferred intentions that follow from the intentions that were explicitly agreed upon. Only the intentions that contribute to Matched-pair are counted in the final solution, and the utility of these intentions is the sum of the utilities of the two furniture items, plus the utility of the Matched-pair (50 points).

Zero-invalids task

The Zero-invalids task is fault intolerant (see Fig. 9). The assumption in the Zero-invalids task is that any mistake invalidates the whole plan. This is a feature of task determinacy: while there are still many possible 8 step plans, all of the solutions with less than 8 items that would be counted as valid solutions for the Standard task are not valid solutions for the Zero-invalids task.

This task is an example of one extreme of fault intolerance. In general, how fault tolerant a task is depends on the interdependency of different subparts of the problem solution. For some tasks, a mistake can invalidate the whole solution, for other tasks, partial solutions without the invalid step may be adequate. For example, in a task like furnishing a room it may be desirable to have both a couch and a chair, but if the agents make a mistake and assume they can use a chair that will end up in a different room, the room is still partially furnished and usable. On the other hand, in a task such as building a tower, each step depends on the successful execution of the previous step and the whole plan may be invalid if a step to put down a foundation block cannot be executed.

Note that an agent can reject another agent's proposal based on believing that it would add an invalid step to the plan, as shown in the rejection utterance act schema in Section 3.2. Since agents have to agree on each step of the plan, an invalid step can only be inserted into the plan if both agents have failed to remember that the preconditions for the plan do not hold.

4.5. *Varying communicative strategies*

Section 3 discussed the discourse act schema that controls how agents participate in dialogue, and discussed the types of utterance acts that the discourse acts are composed of. Which utterance acts a discourse act decomposes into depends on COMMUNICATIVE STRATEGIES which codify different communicative choices for how to do a particular discourse action. Agents are parameterized for different communicative strategies by placing different expansions of discourse plans in their plan libraries.

Varying an agent's communicative strategies provides the basis for testing the hypotheses about the potential benefits of IRUs. Varying the degree of explicitness of a discourse act is the basis of the four communicative strategies tested below: (1) All-implicit; (2) Close-consequence; (3) Explicit-warrant; and (4) Matched-pair-inference-explicit. All of these strategies are hypothesized to mitigate agents' attentional and inferential resource limits, under the assumptions about their architecture and the definition of quality of solution for the task.

Figs. 10–13 show a plan operator for each strategy. These operators draw on work by Walker and Rambow [138], and also make use of Moser and Moore's definitions of discourse acts and the integration of rhetorical structure theory (RST) and Grosz and Sidner's theory of discourse [50,89,93,94]. The predicates in the plan operators are precisely defined by the collaborative planning model and agent architecture discussed in Section 3. Each discourse act, such as a proposal, is composed of a CORE act which represents the primary purpose of the act such as a propose utterance act, as well as a CONTRIBUTOR act, such as a warrant, whose purpose is to increase the likelihood of achieving the intention of the core [92,145,146].

All-implicit strategy

The All-implicit strategy is an expansion of a discourse plan to make a PROPOSAL, in which a PROPOSAL decomposes trivially to the communicative act of PROPOSE. See the plan operator in Fig. 10. This strategy is the communicative choice shown in (3)

NAME:	Proposal-All-implicit (?speaker, ?hearer, ?act)
EFFECT:	(desire ?hearer (do ?hearer ?act) ?utility-act)
CONSTRAINTS:	(and (option ?act)
	(salient ?hearer (utility ?act ?utility-act)))
CORE:	(propose ?speaker ?hearer ?act)

Fig. 10. The PROPOSAL plan operator for an All-implicit agent.

in Section 1, and provides a baseline strategy that is consistent with the REDUNDANCY CONSTRAINT. The experiments below will compare the performance of agents using the All-implicit strategy with the performance of agents using the other proposal strategies discussed below.

In dialogue (12) in Section 4.1, both Design-World agents communicate using the All-implicit strategy, and the proposals are shown in utterances (1), (2), and (3). As Fig. 10 shows, the All-implicit strategy includes no additional information in proposals, leaving it up to the other agent to retrieve them from memory.

The constraints on using the All-implicit strategy are that (1) the proposed ?act is an OPTION generated by means-end reasoning and (2) that the utility is SALIENT to the hearer. In the experiments below, agents are parameterized to use this strategy consistently, so that an agent using the All-implicit strategy assumes everything the hearer knows is always salient. The effect of the proposal is that the hearer will evaluate that proposal and deliberate the degree to which the hearer desires the act. However, whether the hearer will accept or reject the proposal depends on other options the hearer knows about. Clearly the speaker cannot predict these other options. Thus the effect of the proposal does not specify that the action will be intended by the hearer. This holds for all proposal operators.

The All-implicit strategy can be used by agents in any of the Design-World tasks discussed in Section 4.4, since the agents are capable of making inferences or accessing memory to fill in what has been left implicit with this strategy. Other inferences drawn by the hearer from the proposal utterance act are not shown in Fig. 10. For example, an agent can use the All-implicit strategy in either of the Matched-pair tasks, leaving it up to the other agent to infer which other intention makes a match with the option currently under consideration.

Close-consequence

In dialogue (14), agent CLC uses the Close-consequence strategy. The plan operator for this strategy is shown in Fig. 11. The core of the strategy is explicit CLOSING statements, such as (14.2), on the completion of the intention associated with a discourse segment. A contributor to CLC's CLOSING discourse act is an IRU such as (14.3): CLC makes the inference explicit that since they have agreed on putting the green rug in the study, they no longer have the green rug (act-effect inference).

(14) (1) BILL: *Then, let's put the green rug in the study.*
(Propose agent-bill agent-clc option-30:
put-act (agent-bill green rug room-1))

NAME:	Close-consequence (?speaker,?hearer,?act)
EFFECT:	(and (salient ?hearer (effect ?act ?effect)) (bel ?hearer (closed-segment ?act)))
CONSTRAINTS:	(and (intend ?speaker ?hearer ?act) (open-segment ?act))
CONTRIBUTOR:	(say ?speaker ?hearer (effect ?act ?effect))
CORE:	(close ?speaker ?hearer ?act)

Fig. 11. The CLOSING plan operator for a Close-consequence agent.

(2) CLC: *So, we've agreed to put the green rug in the study.*
(Close agent-clc agent-bill intended-30:
put-act (agent-bill green rug room-1))
(3) CLC: AND WE NO LONGER HAVE GREEN RUG.
(Say agent-clc agent-bill bel-48: hasn't (agent-bill green rug))

The Close-consequence strategy of making inferences explicit at the close of a segment models the naturally occurring example in Section 2.2. In both cases an inference is made explicit that follows from what has just been said, and the inference is sequentially located at the close of a discourse segment. This strategy can be used by agents in any of the Design-World tasks discussed in Section 4.4.

The Close-consequence strategy will be used to test hypothesis C2 about potential benefits of making inferences explicit, and will be contrasted with the All-implicit strategy where no closing acts are produced. Note in dialogue (12) in Section 4.2 that both the agents go on to the next phase of the plan, leaving the infererategy will be used to test hypothesis C2 about potential benefits of making inferences explicit, and will be contrasted with the All-implicit strategy where no closing acts are produced. Note in dialogue (12) in Section 4.2 that both the agents go on to the next phase of the plan, leaving the inference of both acceptance and closing for the other agent to make. However, Close-consequence is not a good test of other hypotheses because in the experiments both agents always make act-effect inferences, and these inferences are not difficult to make. See [137] for a discussion of experiments which vary an agent's capability to make these inferences.

Explicit-warrant

The Explicit-warrant strategy varies the proposal discourse act by including WARRANT IRUs in each proposal. The plan operator is given in Fig. 12 and exemplified by the dialogue excerpt in (15). Remember that a WARRANT for an intention is a reason for adopting the intention, and here WARRANTS are the score propositions that give the utility of the proposal, which are mutually believed at the outset of the dialogues. In (15), the WARRANT IRU in (15.1) contributes to the proposal (core act) in (15.2).

(15) (1) IEI: PUTTING IN THE GREEN RUG IS WORTH 56.
(Say agent-iei agent-iei2 bel-2:
score (option-2: put-act (agent-iei green rug room-1) 56))

NAME:	Proposal-Explicit-warrant (?speaker, ?hearer, ?act)
EFFECT:	(and (desire ?hearer (do ?hearer ?act) ?utility-act) (salient ?hearer (utility ?act ?utility-act)))
CONSTRAINTS:	(and (option ?act) (not-salient ?hearer (utility ?act ?utility-act)))
CONTRIBUTOR:	(say ?speaker ?hearer (utility ?act ?utility-act))
CORE:	(propose ?speaker ?hearer ?act)

Fig. 12. The PROPOSAL plan operator for an Explicit-warrant agent.

(2) IEI: *Then, let's put the green rug in the study.*
(Propose agent-iei agent-iei2 option-2:
put-act (agent-iei green rug room-1))

The plan operator in Fig. 12 specifies that an effect of using this plan is that the utility of the proposal option is salient. Since warrants are used by the other agent in deliberation, the Explicit-warrant strategy can save the other agent the processing involved with determining which facts are relevant for deliberation and retrieving them from memory. A constraint on using the Explicit-warrant plan operator is that the utility of the proposal act is not already salient.

In the experiments below, agents are parameterized to use this strategy consistently, with the result that an agent using the Explicit-warrant strategy assumes that the warrant is never salient for the hearer. See [64] for experiments in which an agent attempts to maintain a dynamic model of what is salient for the other agent. The Explicit-warrant strategy also occurs in natural dialogues as shown in the naturally occurring example in dialogue (2).

This strategy can be used by agents in any of the Design-World tasks discussed in Section 4.4. The Explicit-warrant strategy provides a test of hypothesis A1: agents produce Attention IRUs to support the processes of deliberating beliefs and intentions. It can also be used to test hypothesis A4: the choice to produce an Attention IRU is related to the degree to which an agent is resource limited in attentional capacity. In the Standard task this is predicted to improve the performance of resource limited agents. In the Zero-non-matching beliefs task this strategy should increase the likelihood that agents coordinate their beliefs about the warrants underlying different plan steps.

The Matched-pair-inference-explicit strategy

The Matched-pair-inference-explicit strategy expands the PROPOSAL discourse act to two communicative acts. See Fig. 13. The contributor of the proposal consists of statement about what is already intended, while the core is a propose utterance act, as in (16.6) followed by (16.7) in one turn:[16]

[16] The names of agents who use the Matched-pair-inference-explicit strategy are a numbered version of the string "IMI" which stands for Implicit acceptance, matched inference.

NAME:	Proposal-matched-pair (?speaker, ?hearer, ?act1)
EFFECT:	(and (desire ?hearer (do ?hearer ?act1) ?utility-act)
	(salient ?hearer (intend ?speaker ?hearer ?act2)))
CONSTRAINTS:	(and (option ?act1)
	(matched-pair ?act1 ?act2)
	(salient ?hearer (utility ?act1 ?utility-act)))
	(not (salient ?hearer (intend ?speaker ?hearer ?act2)))
CONTRIBUTOR:	(say ?speaker ?hearer (intend ?speaker ?hearer ?act2))
CORE:	(propose ?speaker ?hearer ?act1)

Fig. 13. The PROPOSAL plan operator for a Matched-pair-inference-explicit agent.

(16) (6) IMI2: WE AGREED TO PUT THE PURPLE COUCH IN THE STUDY.
(Say agent-imi2 agent-imi intended-51:
put-act (agent-imi2 purple couch room-1))
(7) IMI2: *Then, let's put the purple rug in the living room.*
(Propose agent-imi2 agent-imi option-80:
put-act (agent-imi2 purple rug room-2))

The statement in (16.6) is an IRU, because it realizes information previously inferred by both agents, and models the IRU in dialogue (11). Matched-pair-inference-explicit is the variation discussed in Section 1 in choice (4). This strategy is only intended to be used in the Matched-pair tasks as a way of testing hypothesis A2, the discourse inference constraint. As Fig. 13 shows, a constraint on using this plan is that the speaking agent has inferred a Matched-pair for the option being proposed.

Although this strategy is specifically tied to Matched-pair inferences, it provides a test of a general strategy for making premises for inferences salient, in tasks that are inferentially complex, and which also require agents to remain coordinated on inferences. For example, to generalize this strategy to other cases of plan-related inferences, the clauses for (Matched-pair ?act1 ?act2) could be replaced with the more general (Generates ?act1 $\wedge$?act2 ?act3), where the generates relation is to be inferred [33,97].

Note that the effect of using this strategy is not that the hearer makes the Matched-pair inference, rather the effect is that the premise for the desired inference is salient. A constraint on using this strategy is that this premise is not already salient. However, agents parameterized with this strategy always assume that the premise is not salient for the hearer. See [64] for experiments in which agents attempt to maintain a dynamic model of the other agent's attentional state.

4.6. *Plan evaluation*

Section 3.3 specified a model of how collaborative plans are evaluated in terms of QUALITY OF SOLUTION and COLLABORATIVE EFFORT. Design-World is constructed in order to be able to measure the quality of a solution as well as collaborative effort. Section 4.4 defined quality of solution for all of the Design-World tasks. We want to examine trade-offs in performance between strategy choices.

It is obvious that these trade-offs can be related to the relative contributions of total cost of communication versus the total cost of inference versus the total cost of retrieval for both agents' collaborative effort. Thus, to calculate collaborative effort, we cannot simply add up the number of retrievals, inferences and messages. Consider that a Consequence IRU that makes an inference explicit ensures that an inferred belief becomes part of the discourse model. However, if the inference would have been made anyway, the benefit of the strategy is dependent upon whether the effort to make the inference without the Consequence IRU would have been greater than the cost of processing the extra utterance of the Consequence IRU. A similar argument holds for the potential benefit of Attention IRUs. Whenever an Attention IRU reduces overall effort for retrieval while not increasing communication effort to the same degree, it will be beneficial. This hypothesis is given below in a general form.

- HYPOTH-I1: Strategies that reduce collaborative effort without affecting quality of solution are beneficial.

This hypothesis follows directly from the definition of performance repeated here for convenience from Section 3.3:

$$\text{PERFORMANCE} = \text{QUALITY OF SOLUTION} - \text{COLLABORATIVE EFFORT}.$$

We need to introduce parameters for the effort involved with each of the component processes because they are not strictly comparable, and because these modules are implementation dependent. Thus agents' retrieval, inference and communicative costs are parameterized by (1) COMMCOST: cost of sending a message; (2) INFCOST: cost of inference; and (3) RETCOST: cost of retrieval from memory. Collaborative effort is then defined as:

$$\begin{aligned}\text{COLLABORATIVE EFFORT} = {} & (\text{COMMCOST} \times \text{total messages for both agents})\\ & +(\text{INFCOST} \times \text{total inferences for both agents})\\ & +(\text{RETCOST} \times \text{total retrievals for both agents}).\end{aligned}$$

We will use these cost parameters to explore three extremes in this space: (1) when processing is free; (2) when retrieval effort dominates other processing costs; and (3) when communication effort dominates other processing costs. The parameters support modeling various instantiations of the agent architecture given in Fig. 3. For example, varying the cost of retrieval models different assumptions about how the beliefs database, plan library and working memory are implemented. Varying the cost of communication models situations in which communication planning is very costly. The relation between the values of these parameters and the utilities of the steps in the plan determines experimental outcomes, rather than the absolute values.

As an example of the effect of varying these costs, consider the plots of performance distributions shown in Figs. 14 and 15 for LOW, MID and HIGH AWM. In these figures, performance is plotted on the x-axis and number of simulations at that performance level are given by bars on the y-axis. The performance distributions in Fig. 14 demonstrate the increase in QUALITY OF SOLUTION that we would expect with increases in AWM,

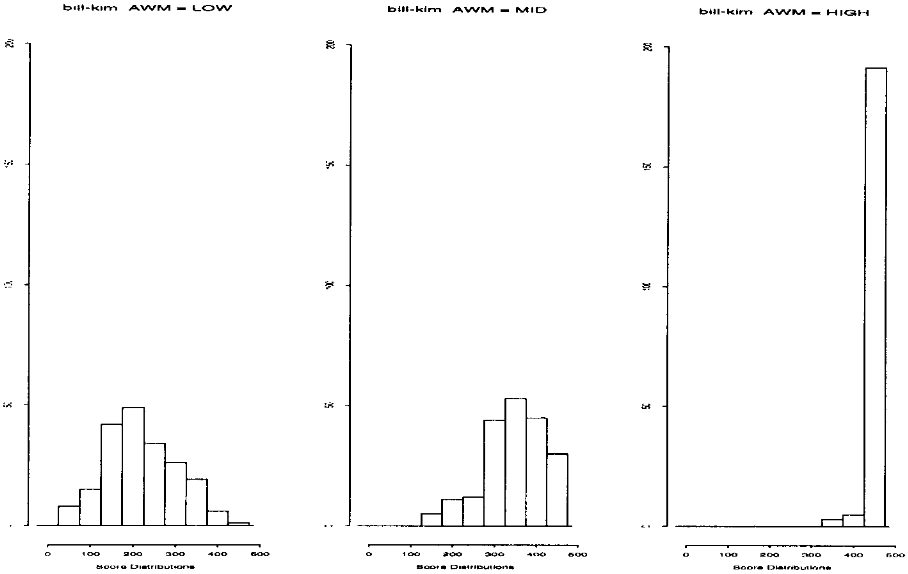

Fig. 14. Performance distributions showing the effect of AWM parameterization for dialogues between two All-implicit agents when all processing is free. The three performance distributions are for LOW, MID and HIGH AWM agents.

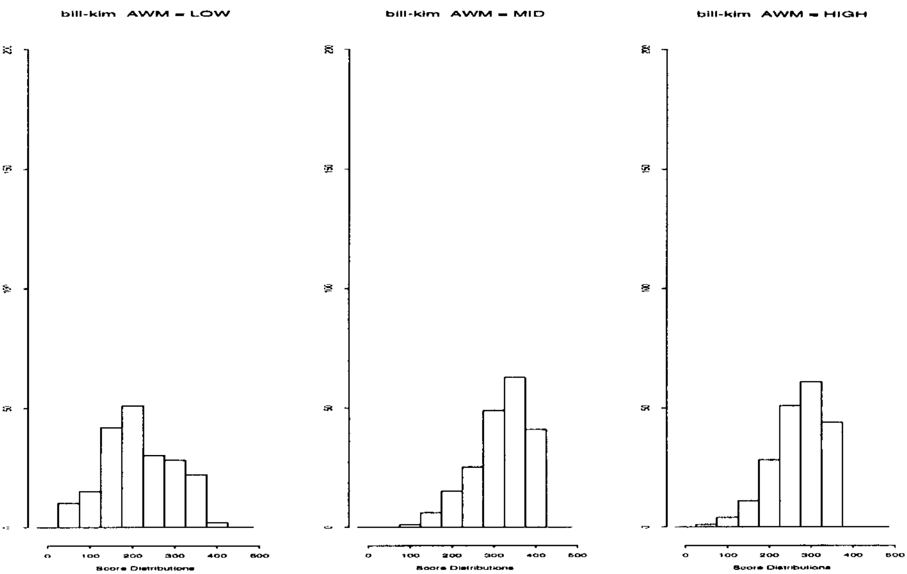

Fig. 15. Performance distributions showing the effect of increased retrieval cost for each AWM range for dialogues between two All-implicit agents. The three performance distributions are for LOW, MID and HIGH AWM agents. retcost = 0.001.

given no processing costs.[17] Fig. 15 shows what happens when processing is not free: here a retrieval cost of 0.001 means that every memory access reduces quality of solution by 1/1000 of a point (remember that the utilities of plan steps range between 10 and 56). As Fig. 15 shows, the ability to access the whole beliefs database in reasoning does not always improve performance since HIGH AWM agents perform similarly to MID AWM agents.

4.7. Summary: mapping from naturally occurring data to Design-World experiments

Section 2 proposed hypotheses about the function of IRUs in human to human collaborative planning dialogues, and then Section 3 presented a model for collaborative planning dialogues based on the observations in Section 2. Section 4 then described Design-World as a testbed of the model, and Sections 4.4 and 4.5 introduced a number of parameters of the testbed that are intended to model the features of the human-human dialogues and support testing of the hypotheses. Here I wish to summarize the mapping between the naturally occurring dialogues and the design of the testbed in order to clarify the basis for the experiments in the next section.

The testbed and the experimental parameters are based on the following mapping between human-human collaborative planning dialogues and the testbed. First, the planning and deliberation aspects of human processing are modeled with the IRMA architecture, and resource limits on these processes are modeled by extending the IRMA architecture with a model of attention/working memory (AWM) which has been shown to model a limited but critical set of properties of human processing. Second, the processing of dialogue is tied to the agent architecture. Third, the mapping of a WARRANT relation between an act and a belief in naturally occurring examples such as (9) is modeled with a WARRANT relation between an act and a belief in Design-World as seen in the Explicit-warrant communication strategy in Section 4.5. Fourth, the mapping assumes that arbitrary content-based inferences in natural dialogues such as that discussed in relation to example (11) can be mapped to content-based inferences in Design-World such as those required for doing well on the Matched-pair tasks. Fifth, the mapping is based on the assumption that task difficulty in naturally occurring tasks such as those in the financial advice domain can be related to three abstract features: (1) inferential complexity as measured by the number of premise required for making an inferences; (2) degree of belief coordination required on intentions, inferences and beliefs underlying a plan; and (3) task determinacy and fault tolerance. Finally the mapping assumes that it is reasonable to evaluate the performance of the agents in collaborative planning dialogues by using domain plan utility for a measure of the quality of solution and defining the cost to achieve that solution as collaborative effort, appropriately parameterized.

---

[17] These distributions approximate Beta distributions [144], and this approximation was used to determine that 200 runs would guarantee stable results. The Beta distribution with the largest variance, for parameters R and S greater than or equal to 1, is the uniform distribution. This largest variance distribution would require approximately 133 samples [115, 144]. An empirical evaluation of the adequacy of this sample size for three different strategies was tested to see if any differences showed up in alternate runs of 100; no differences were found.

The details of this mapping specifies how the testbed implements the model of collaborative planning and provides the basis for extrapolating from the testbed experimental results to the human-human dialogues that are being modeled. The testbed provides an excellent environment for testing the hypotheses to the extent that the model captures critical aspects of human-human dialogues.

5. Experimental results

5.1. Statistically evaluating performance

The experiments examine the interaction between tasks, communication strategies and AWM resource limits. Every experiment varies AWM over three ranges: LOW, MID, and HIGH. In order to run an experiment on a particular communicative strategy for a particular task, 200 dialogues for each AWM range are simulated. Because the AWM model is probabilistic, each dialogue simulation has a different result. The AWM parameter yields a performance distribution for very resource limited agents (LOW), agents hypothesized to be similar to human agents (MID), and resource unlimited agents (HIGH). Sample performance distributions for QUALITY OF SOLUTION (with no collaborative effort subtracted) from runs of two All-implicit agents for each AWM setting are shown in Fig. 14.

To test our hypotheses, we want to *compare* the performance of two different communicative strategies for a particular task, under different asssumptions about resource limits and processing costs. To see the effect of communicative strategy and AWM over the whole range of AWM settings, we first run a two-way analysis of variance (anova) with AWM as one factor and communication strategy as another. [18] The anova tells us whether: (1) AWM alone is a significant factor in predicting performance; (2) communication strategy alone is a significant factor in predicting performance; and (3) whether there is an interaction between communication strategy and AWM.

However, anova alone does not enable us to determine the particular AWM range at which a communication strategy aids or hinders performance, and many of the hypotheses about the benefits of particular communication strategies are specific to how resource limited an agent is. Furthermore, whenever strategy affects performance positively for one value of AWM and negatively for another value of AWM, the potential effects of strategy cannot be seen from the anova alone. Therefore, we conduct planned comparisons of strategies using the modified Bonferroni test (hereafter MB) [70], within each AWM range setting to determine which AWM range the strategy affects [25,70]. [19] On the basis of these comparisons we can say whether a strategy is BENEFICIAL for a particular task for a particular AWM range.

[18] The experimental performance distributions are not normal and the variance is not the same over different samples, however anova is robust against the violation of these assumptions under the conditions in these experiments [25,70].

[19] According to the modified Bonferroni test, the significant F values for the planned comparisons reported below are 3.88 for $p < 0.05$, 5.06 for $p < 0.025$, 6.66 for $p < 0.01$, and 9.61 for $p < 0.002$.

> A strategy A is BENEFICIAL as compared to a strategy B, for a particular AWM range, in the same task situation, with the same cost settings, if the mean of A is significantly greater than the mean of B, according to the modified Bonferroni test (MB) test.

The converse of BENEFICIAL is DETRIMENTAL:

> A strategy A is DETRIMENTAL as compared to a strategy B, for a particular AWM range, in the same task situation, with the same cost settings, if the mean of A is significantly less than the mean of B, according to the modified Bonferroni test (MB) test.

Strategies need not be either BENEFICIAL or DETRIMENTAL, there may be no difference between two strategies. Also with the definition given above a strategy may be both BENEFICIAL and DETRIMENTAL depending on the range of AWM that the two strategies are compared over, i.e., a strategy may be beneficial for LOW AWM agents and detrimental for HIGH AWM agents.

A DIFFERENCE PLOT such as that in Fig. 16 is used to summarize a comparison of two strategies, strategy 1 and strategy 2. In the comparisons below, strategy 1 is either Close-consequence,[20] Explicit-warrant, or Matched-pair-inference-explicit and strategy 2 is the All-implicit strategy. *Differences* in performance means between two strategies are plotted on the y-axis against AWM ranges on the x-axis. Each point in the plot represents the difference in the means of 200 runs of each strategy at a particular AWM range. These plots summarize the information from 1200 simulated dialogues.

5.2. *Standard task*

Remember that the Standard task is defined so that the QUALITY OF SOLUTION that agents achieve for a DESIGN-HOUSE plan, constructed via the dialogue, is the sum of the utilities of each valid step in their plan. The task has multiple correct solutions and is fault tolerant because the point values for invalid steps in the plan are simply subtracted from the score, with the effect that agents are not heavily penalized for making mistakes. Furthermore, the task has low inferential complexity: the only inferences agents are required to make are those for deliberation and means-end reasoning. In both of these cases, to make these inferences, agents are only required to access a single minor premise.

All-implicit agents do fairly well at the Standard task, under assumptions that all processing is free, as shown in the performance plot in Fig. 14. However, as retrieval costs increase, HIGH AWM agents don't do as well as when retrieval is free, because they expend too much effort on retrieval during collaborative planning. Compare the HIGH AWM distribution in Fig. 14 with that in Fig. 15. Thus for the Standard task, HIGH AWM agents have the potential to benefit from communication strategies that

[20] In experiments with Close-consequence only one agent in a dialogue uses the Close-consequence strategy because the use of this strategy is constrained to when the dialogue segment is open. See Fig. 11. Since only one agent will ever produce a closing statement for any dialogue segment, only one agent is given the option in the simulations.

reduce the total effort for retrieval, when retrieval is not free. In addition, although the task has minimal inferential complexity, easy access to information that is used for deliberation, which the Explicit-warrant strategy provides, could benefit LOW AWM agents, since they might otherwise make non-optimal decisions. Furthermore, although the task is fault tolerant, agents are still penalized for making errors since errors do not contribute to performance. Thus for the Standard task, communication strategies such as Close-consequence that can reduce the number of errors could be beneficial. Below we will compare the All-implicit strategy to the Explicit-warrant strategy and the Close-consequence strategy.

Explicit-warrant

The Explicit-warrant strategy can be used in the Standard task to test hypothesis A1: agents produce Attention IRUs to support the processes of deliberating beliefs and intentions. It can also be used to test hypothesis A4: the choice to produce an Attention IRU is related to the degree to which an agent is resource limited in attentional capacity. Thus one prediction is that the Explicit-warrant strategy will result in higher performance for LOW AWM agents even when processing is free by ensuring that they can access the warrant and use it in deliberation, thus making better decisions.

Fig. 16 plots the differences in the performance means between the Explicit-warrant strategy and the All-implicit strategy for LOW, MID and HIGH AWM agents. A two-way anova exploring the effect of AWM and the Explicit-warrant strategy for the Standard task shows that AWM has a large effect on performance ($F = 336.63$, $p < 0.000001$). There is no main effect for communicative strategy ($F = 1.92$, $p < 0.16$). However, there is an interaction between AWM and communicative choice ($F = 1136.34$, $p < 0.000001$).

By comparing performance within a particular AWM range for each strategy we can see which AWM settings interact with communicative strategy. The planned comparisons using the modified Bonferonni (MB) test show that the Explicit-warrant strategy is neither beneficial nor detrimental in the Standard task, in comparison with the All-implicit strategy, if all processing is free (MB(LOW) = 0.29, ns; MB(MID) = 2.79, ns; MB(HIGH) = 0.39, ns). Note that there is a trend towards the Explicit-warrant strategy being detrimental at MID AWM.

The hypothesis based on the corpus analysis was that LOW AWM agents might benefit from communicative strategies that include IRUs. However, this hypothesis is disconfirmed. Further analysis of this result suggests a hypothesis not apparent from the corpus analysis: any beneficial effect of an IRU can be cancelled for resource limited agents because IRUs may displace other information from working memory that is more useful. In this case, despite the fact that the warrant information is useful for deliberation, making the warrant salient displaces information that can be used to generate other options. When agents are very resource limited making an optimal decision is not as important as being able to generate multiple options.

The Explicit-warrant strategy can also be used in the Standard task to test hypothesis I1: strategies that reduce collaborative effort overall may be beneficial. Thus, another prediction is that by providing the warrant used in deliberating a proposal with every proposal, the Explicit-warrant strategy has the potential to reduce resource consumption when accessing memory has some processing cost.

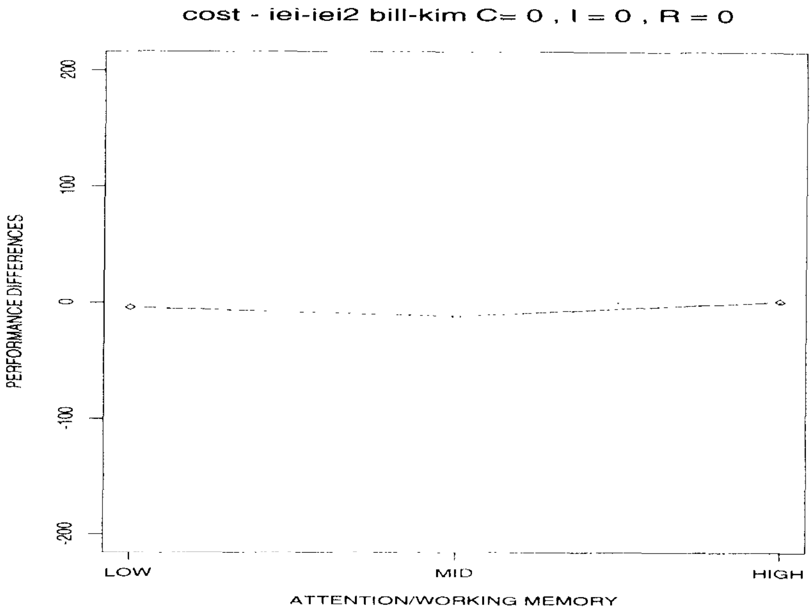

Fig. 16. If processing is free, Explicit-warrant is neither beneficial nor detrimental for all AWM settings: strategy 1 of two Explicit-warrant agents and strategy 2 of two All-implicit agents: task = Standard, commcost = 0, infcost = 0, retcost = 0.

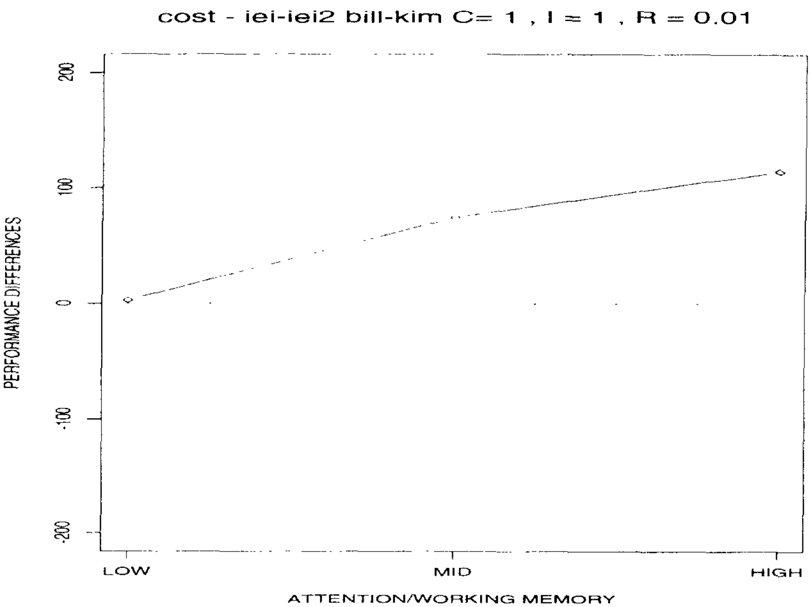

Fig. 17. Explicit-warrant is beneficial for MID and HIGH AWM agents when retrieval dominates processing costs: strategy 1 is two Explicit-warrant agents and strategy 2 is two All-implicit agents: task = Standard, commcost = 1, infcost = 1, retcost = 0.01.

Fig. 17 plots the differences in the performance means between the Explicit-warrant strategy and the All-implicit strategy for LOW, MID and HIGH AWM agents when retrieval effort dominates processing. A two-way anova exploring the effect of AWM and the Explicit-warrant strategy for the Standard task, when retrieval cost dominates processing, shows that AWM has a large effect on performance ($F = 330.15$, $p < 0.000001$). There is also a main effect for communicative strategy ($F = 5.74$, $p < 0.01$), and an interaction between AWM and communicative choice ($F = 1077.64$, $p < 0.000001$).

The planned comparisons using the MB test to compare performance at each AWM range show that, in the Standard task, in comparison with the All-implicit strategy, the Explicit-warrant strategy is neither beneficial nor detrimental for LOW AWM agents (MB(LOW) = 0.27, ns). However, hypothesis I1 is confirmed because the Explicit-warrant strategy is beneficial for MID AWM agents (MB(MID) = 86.43, $p < 0.002$). The Explicit-warrant strategy also tends towards improving performance for HIGH AWM agents (MB(HIGH) = 2.07, $p < 0.10$). For higher AWM values, this trend is because the beliefs necessary for deliberating the proposal are made available in the current context with each proposal, so that agents don't have to search memory for them.

As an additional test of hypothesis I1, a final experiment tests the Explicit-warrant strategy against the All-implicit strategy in a situation where the cost of communication dominates other processing costs. Fig. 18 plots the differences in the performance means between the Explicit-warrant strategy and the All-implicit strategy for LOW, MID and HIGH AWM agents when communication effort dominates processing. A two-way anova exploring the effect of AWM and the Explicit-warrant strategy for the Standard task, when communication effort dominates processing, shows that AWM has a large effect on performance ($F = 409.52$, $p < 0.000001$). There is also a main effect for communicative strategy ($F = 28.12$, $p < 0.000001$), and an interaction between AWM and communicative choice ($F = 960.24$, $p < 0.000001$).

The planned comparisons using the MB test to compare performance at each AWM range show that in this situation, when communication effort dominates processing, the Explicit-warrant strategy is neither beneficial nor detrimental for MID AWM agents (MB(MID) = 0.12, ns). However, the Explicit-warrant strategy is detrimental for both LOW and HIGH AWM agents (MB(LOW) = 7.69, $p < 0.01$; MB(HIGH) = 39.65, $p < 0.01$). Since this strategy includes an extra utterance with every proposal and provides no clear benefits, it is detrimental to performance in the Standard task when communication effort dominates processing. Below, when we compare this situation with that in the Zero-non-matching-beliefs task, we will see that this is due to the fact that the Standard task has low coordination requirements.

Close-consequence

The Close-consequence strategy of making inferences explicit can be used in the Standard task to test hypothesis C4: the choice to produce a Consequence IRU is related to a measure of "how important" the inference is. Even though the Standard task is fault tolerant, every invalid step reduces the quality of solution of the final plan. Making act-effect inferences explicit decreases the likelihood of making this kind of error.

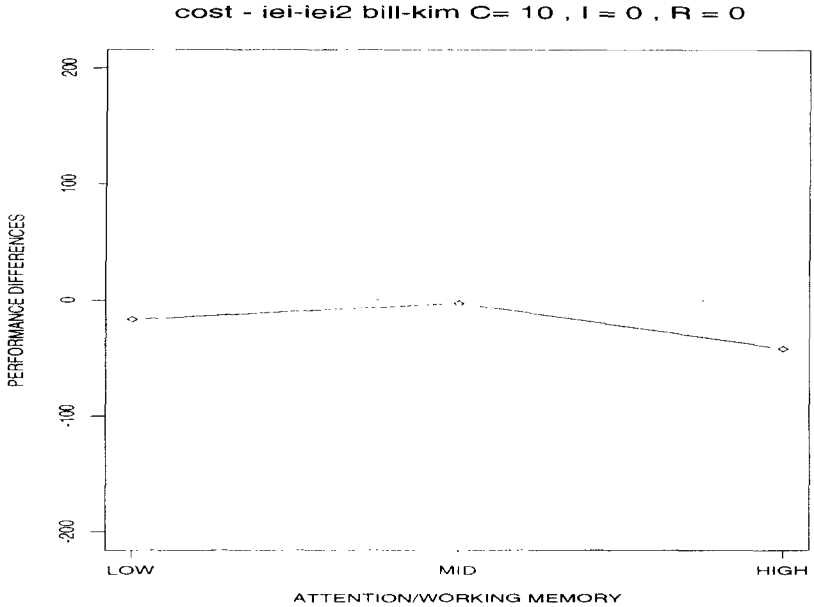

Fig. 18. Explicit-warrant is detrimental for LOW and HIGH AWM agents when communication effort is high: strategy 1 is two Explicit-warrant agents and strategy 2 is two All-implicit agents: task = Standard, commcost = 10, infcost = 0, retcost = 0.

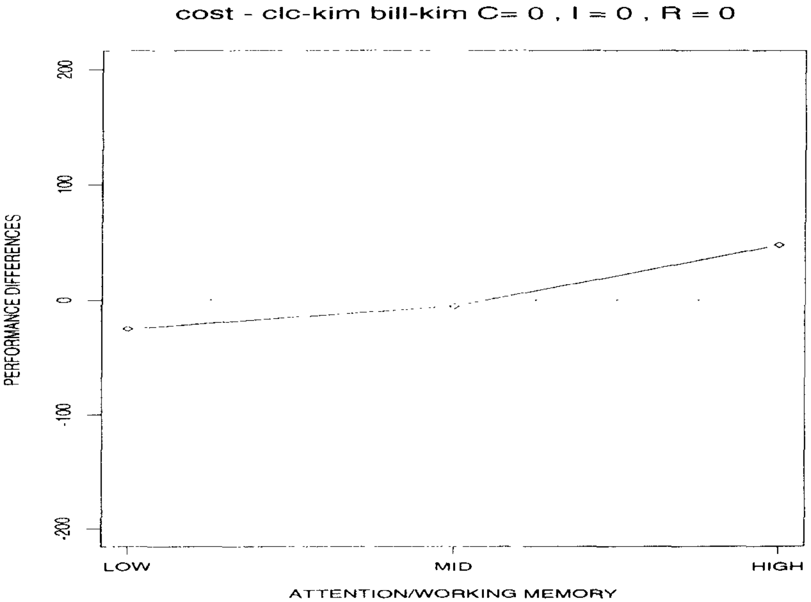

Fig. 19. Close-consequence can be detrimental in the Standard task for LOW AWM agents and beneficial for HIGH AWM agents. Strategy 1 is the combination of an All-implicit agent with a Close-consequence agent and strategy 2 is two All-implicit agents, task = Standard, commcost = 0, infcost = 0, retcost = 0.

The difference plot in Fig. 19 plots performance differences between the Close-consequence strategy and the All-implicit strategy, in the Standard task, when all processing is free. A two-way anova exploring the effect of AWM and the Close-consequence strategy in this situation shows that AWM has a large effect on performance ($F = 249.20$, $p < 0.000001$), and that there is an interaction between AWM and communicative choice ($F = 919.27$, $p < 0.000001$).

Planned comparisons between strategies for each AWM range shows that the Close-consequence strategy is detrimental in comparison with All-implicit for LOW AWM agents (MB(LOW) = 8.70, $p < 0.01$). This is because generating options contributes more to performance for agents with LOW AWM than avoiding errors, and the additional utterances that make inferences explicit in the Close-consequence strategy has the effect of displacing facts that could be used in means–end reasoning to generate options. There is no difference in performance for MID AWM agents (MB(MID) = 0.439, ns).

However, comparisons between the two strategies for HIGH AWM agents shows that the Close-consequence strategy is beneficial in comparison with All-implicit (MB(HIGH) = 171.71, $p < 0.002$). See Fig. 19. This is because the belief deliberation algorithm increases the probability of HIGH AWM agents choosing to believe out of date beliefs about the state of the world. The result is that they are more likely to have invalid steps in their plans. Thus the Close-consequence strategy is beneficial because reinforcing the belief that a furniture item has been used makes it less likely that agents will believe that they still have that furniture item. This result is not predicted by any hypotheses, but as discussed in Section 4.3, this property of the belief deliberation mechanism has some intuitive appeal. In any case, this result provides a data point for the benefit of a strategy for making inferences explicit when the probability of making an error increases if that inference is not made.

5.3. Zero-non-matching-beliefs task

Remember that the Zero-non-matching-beliefs task requires a greater degree of belief coordination by requiring agents to agree on the beliefs underlying deliberation (WARRANTS).[21] Thus, it increases the importance of making particular deliberation-based inferences, and can therefore be used to test hypotheses A1, A4 and A5. Below we will compare the performance of agents using the All-implicit strategy with the Explicit-warrant strategy in the Zero-non-matching-beliefs task.

Fig. 20 plots the mean performance differences of the Explicit-warrant strategy and the All-implicit strategy in the Zero-non-matching-beliefs task. A two-way anova exploring the effect of AWM and communicative strategy for the Zero-non-matching-beliefs task, shows that AWM has a large effect on performance ($F = 471.42$, $p < 0.000001$). There is also a main effect for communicative strategy ($F = 379.74$, $p < 0.000001$), and an interaction between AWM and communicative choice ($F = 669.24$, $p < 0.000001$).

[21] Remember that in other tasks, agents do not have to agree on WARRANTS because in situations in which they know of only one option, they do not need to retrieve the warrant in order to be able to decide to accept the proposal. Thus when agents have limited AWM, they may accept a proposal without having retrieved the warrant.

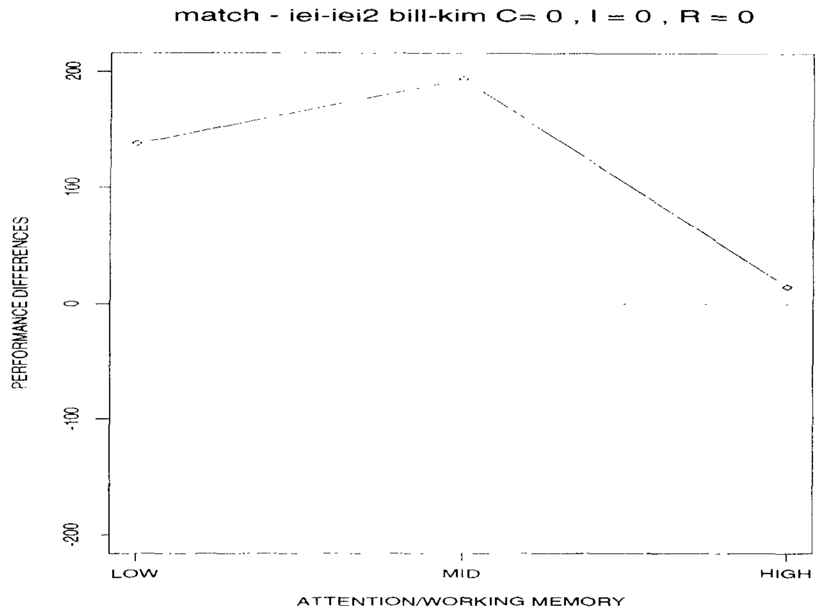

Fig. 20. Explicit-warrant is beneficial for Zero-non-matching-beliefs task for LOW and MID AWM agents: strategy 1 is two Explicit-warrant agents and strategy 2 is two All-implicit agents: task = Zero-non-matching-beliefs, commcost = 0, infcost = 0, retcost = 0.

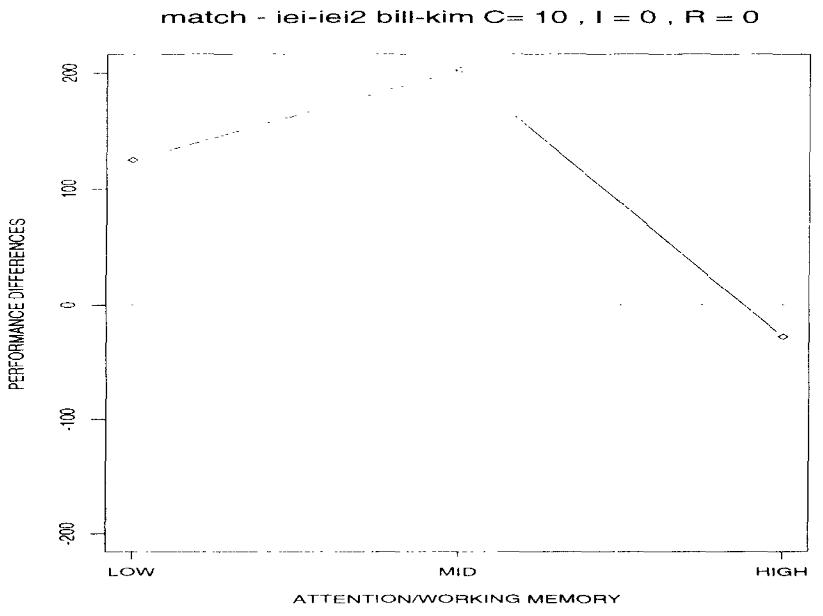

Fig. 21. Explicit-warrant is beneficial for Zero-non-matching-beliefs task, for LOW and MID AWM agents, even when communication cost dominates processing: strategy 1 is two Explicit-warrant agents and strategy 2 is two All-implicit agents: task = Zero-non-matching-beliefs, commcost = 10, infcost = 0, retcost = 0.

Comparisons within each AWM range of the two communicative strategies in this task shows that the Explicit-warrant strategy is highly beneficial for LOW and MID AWM agents (MB(LOW) = 260.6, $p < 0.002$; MB(MID) = 195.5, $p < 0.002$). The strategy is also beneficial for HIGH AWM agents MB(HIGH) = 4.48, $p < 0.05$). When agents are resource limited, they may fail to access a warrant. The Explicit-warrant strategy guarantees that the agents always can access the warrant for the option under discussion. Thus, even agents with higher values of AWM can benefit from this strategy, since the task requires such a high degree of belief coordination.

Hypothesis I1 can also be tested in this task. We can ask whether it is possible to drive the total effort for communication high enough to make it inefficient to choose the Explicit-warrant strategy over All-implicit. However, the benefits of the Explicit-warrant strategy for LOW and MID AWM agents for this task are so strong that they cannot be reduced even when communication cost is high (MB(LOW) = 246.4, $p < 0.002$; MB(MID) = 242.7, $p < 0.002$). See Fig. 21. In other words, even when every extra WARRANT message increases collaborative effort by 10 and reduces performance by 10, if the task is Zero-non-matching-beliefs, resource limited agents using Explicit-warrant do better. Contrast Fig. 21 with the Standard task and same cost parameters in Fig. 18.

However, when communication cost is high, the strategy becomes detrimental for HIGH AWM agents (MB(HIGH) = 7.56, $p < 0.01$). These agents can usually access warrants and the increase in belief coordination afforded by the Explicit-warrant strategy does not offset the high communication cost.

5.4. *Inferential tasks: Matched-pair*

The two versions of the Matched-pair tasks described in Section 4.4 (1) increase the inferential complexity of the task and (2) increase the degree of belief coordination required by requiring agents to be coordinated on inferences that follow from intentions that have been explicitly agreed upon. Both tasks increase inferential difficulty to a small degree: All-implicit agents do fairly well at making Matched-pair inferences. The Matched-pair-same-room task requires the same inferences as the Matched-pair-two-room task, but these inferences should be easier to make in the Matched-pair-same-room since the inferential premises are more likely to be salient.

The Matched-pair tasks provide an environment for testing hypotheses A2, A3, A4 and A5. The Attention strategy that is used to test these hypotheses is the Matched-pair-inference-explicit strategy; this strategy makes the premises for Matched-pair inferences salient, thus increasing the likelihood of agents making Matched-pair inferences. The predictions are that this strategy should be beneficial for LOW and possibly for MID AWM agents, but that HIGH AWM agents can access the necessary inferential premises without Attention IRUs. Furthermore, we predict that the beneficial effect should be stronger for the Matched-pair-two-room task.

Fig. 22 plots the performance differences between All-implicit agents and Matched-pair-inference-explicit agents for the Matched-pair-same-room task. A two-way anova exploring the effect of AWM and communicative strategy in this task, shows that AWM has a large effect on performance ($F = 323.93$, $p < 0.000001$). There is no main effect

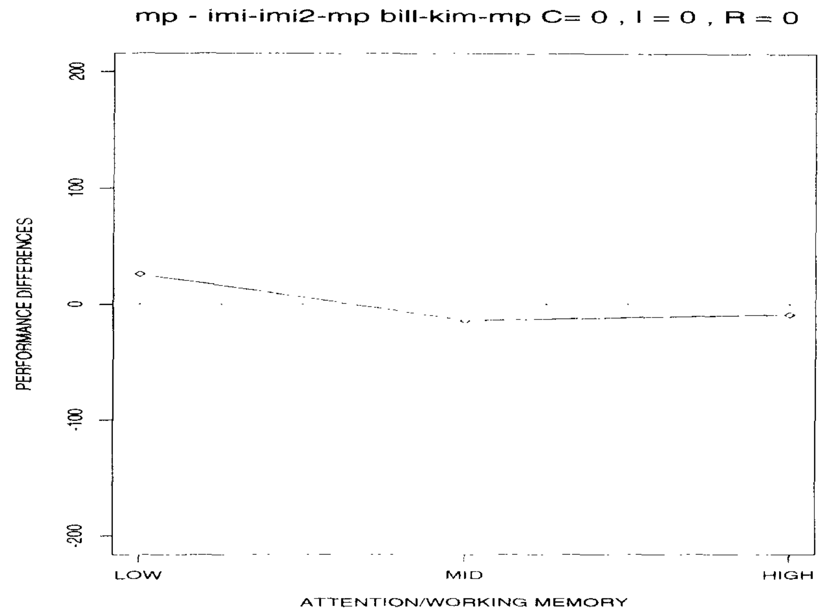

Fig. 22. Matched-pair-inference-explicit is beneficial for LOW AWM agents in Matched-pair-same-room. Strategy 1 is two Matched-pair-inference-explicit agents and strategy 2 is two All-implicit agents, task = Matched-pair-same-room, commcost = 0, infcost = 0, retcost = 0.

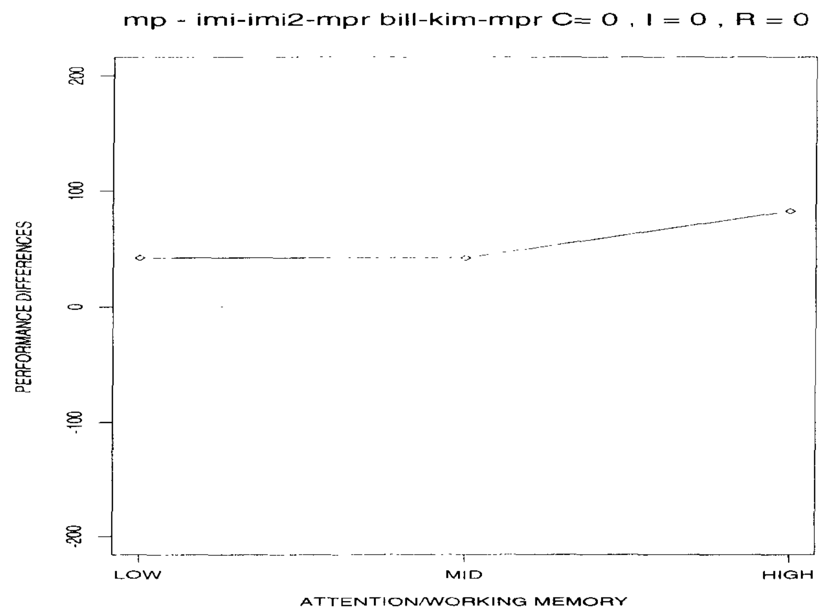

Fig. 23. Matched-pair-inference-explicit is beneficial for LOW, MID and HIGH AWM agents in the Matched-pair-two-room task. Strategy 1 is two Matched-pair-inference-explicit agents and strategy 2 is two All-implicit agents, task = Matched-pair-two-room, commcost = 0, infcost = 0, retcost = 0.

for communicative strategy ($F = 0.03$, ns), but there is an interaction between AWM and communicative choice ($F = 1101.51$, $p < 0.000001$).

Comparisons within AWM ranges between agents using the All-implicit strategy and agents using the Matched-pair-inference-explicit strategy in the Matched-pair-same-room task (Fig. 22) show that Matched-pair-inference-explicit strategy is beneficial for LOW AWM agents (MB(LOW) = 4.47, $p < 0.05$), but not significantly different for either MID or HIGH AWM agents). In the Matched-pair-same-room task the content of the IRU was recently inferred and is likely to still be salient, thus the beneficial effect is relatively small and is restricted to very resource limited agents.

In contrast, in the Matched-pair-two-room task, the effect on performance of the Matched-pair-inference-explicit strategy is much larger, as we predicted. Fig. 23 plots the mean performance differences of agents using the Matched-pair-inference-explicit strategy and those using the All-implicit strategy. The All-implicit agents do not manage to achieve the same levels of mutual inference as Matched-pair-inference-explicit agents. A two-way anova exploring the effect of AWM and communicative strategy in this task, shows that AWM has a large effect on performance ($F = 171.79$, $p < 0.000001$). There is a main effect for communicative strategy ($F = 57.12$, $p < 0.001$), and an interaction between AWM and communicative choice ($F = 567.34$, $p < 0.000001$).

Comparisons within AWM ranges between agents using the All-implicit strategy and agents using the Matched-pair-inference-explicit strategy in the Matched-pair-two-room task (Fig. 23) show that Matched-pair-inference-explicit strategy is beneficial for LOW, MID and HIGH AWM agents (MB(LOW) = 21.94, $p < 0.01$; MB(MID) = 7.71, $p < 0.01$; MB(HIGH) = 38.85, $p < 0.002$). In other words, this strategy is highly effective in increasing the ability of LOW, MID and HIGH AWM agents to make Matched-pair inferences in the Matched-pair-two-room task.

We predicted the strategy to be beneficial for LOW and possibly for MID AWM agents because it gives agents access to premises for inferences which they would otherwise be unable to access. This confirms the effect of the hypothesized DISCOURSE INFERENCE CONSTRAINT. However, we did not expect it to be beneficial for HIGH AWM agents. This surprising effect is due to the fact that, in the case of higher AWM values, the Matched-pair-inference-explicit strategy keeps the agents coordinated on which inference the proposing agent intended in a situation in which multiple inferences are possible. In other words, when agents have HIGH AWM they can make *divergent* inferences, and a strategy of making inferential premises salient improves agents' inferential coordination. Thus the strategy controls inferential processes in a way that was not predicted based on the corpus analysis alone.

Hypothesis I1 can also be tested in this task. We can ask whether it is possible to drive the effort for communication high enough to make it inefficient to choose the Matched-pair-inference-explicit strategy over All-implicit.

Fig. 24 plots the mean performance differences between these two strategies when communication cost is high. Comparisons within each AWM range shows that this strategy is still beneficial for LOW, MID and HIGH AWM agents even with a high communication cost (MB(LOW) = 19.10, $p < 0.01$; MB(MID) = 3.94, $p < 0.05$; MB(HIGH) = 10.46, $p < 0.01$). In other words it would be difficult to find a task situation that required coordinating on inference in which this strategy was not beneficial. This result

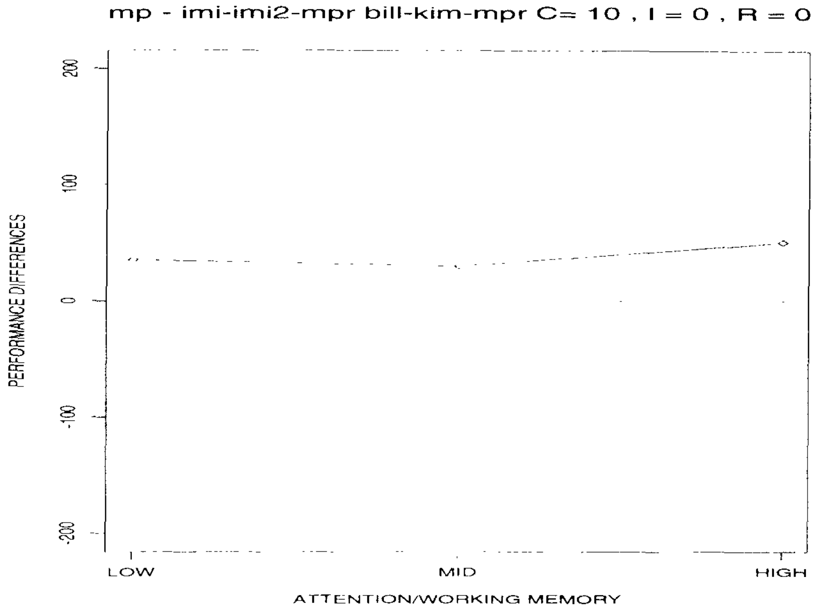

Fig. 24. The Matched-pair-inference-explicit strategy is beneficial for LOW, MID and HIGH AWM agents in the Matched-pair-two-room task even with communication cost of 10. Strategy 1 is two Matched-pair-inference-explicit agents and strategy 2 is two All-implicit agents, task = Matched-pair-two-room, commcost = 10, infcost = 0, retcost = 0.

is strong support for the DISCOURSE INFERENCE CONSTRAINT, which may explain the prevalence of this strategy in naturally occurring dialogues [30, 109, 141].

5.5. *Zero-invalids task*

Remember that the Zero-invalids task is a fault intolerant version of the task in which any invalid intention invalidates the whole plan. Thus the Zero-invalids task provides an environment for testing hypotheses C2 and C4 with respect to the inferences made explicit by the Close-consequence strategy.

Fig. 25 plots the mean performance differences between agents using the Close-consequence strategy and agents using the All-implicit strategy in the Zero-invalids task. A two-way anova exploring the effect of AWM and communicative strategy in this task, shows that AWM has a large effect on performance ($F = 223.14$, $p < 0.000001$). There is a main effect for communicative strategy ($F = 75.81$, $p < 0.001$), and an interaction between AWM and communicative choice ($F = 103.38$, $p < 0.000001$).

The Close-consequence strategy was detrimental in the Standard task for LOW AWM agents. Comparisons within AWM ranges between agents using the All-implicit strategy and agents using the Close-consequence strategy in the Zero-invalids task show that there are no differences in performance for LOW AWM agents in the fault intolerant Zero-invalids task (MB(LOW) = 3.64, ns). However, the Close-consequence strategy is beneficial for MID and HIGH AWM agents (MB(MID) = 26.62, $p < 0.002$; MB(HIGH)

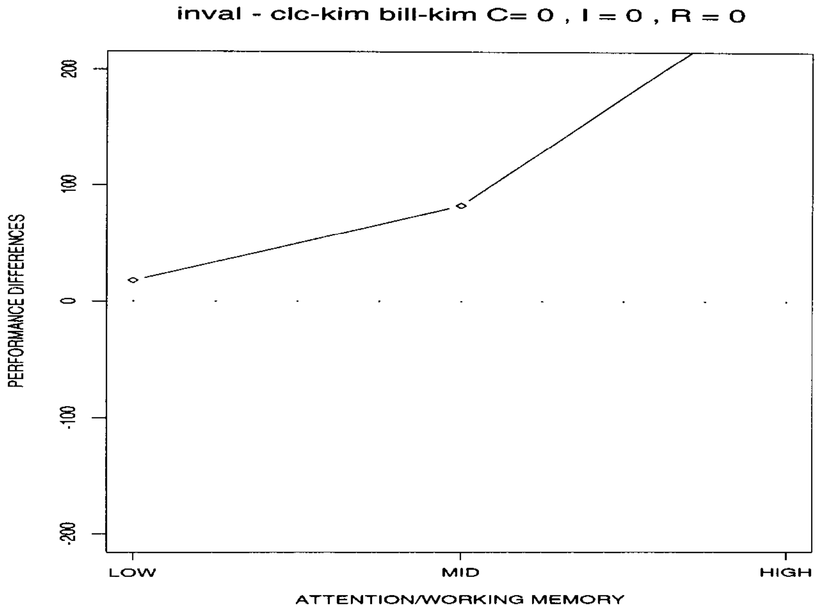

Fig. 25. Close-consequence is beneficial for the Zero-invalids task for MID and HIGH AWM agents. Strategy 1 is the combination of an All-implicit agent with a Close-consequence agent and strategy 2 is two All-implicit agents, task = Zero-invalids, commcost = 0, infcost = 0, retcost = 0.

= 267.72, $p < 0.002$). In other words, this strategy is highly beneficial in increasing the robustness of the planning process by decreasing the frequency with which agents make mistakes. This is a direct result of *rehearsing* the act-effect inferences, making it unlikely that attention limited agents will forget these important inferences.

6. Discussion

This paper showed how agents' choice in communicative action can be designed to mitigate the effect of their resource limits in the context of particular features of a collaborative planning task. In Section 3, I presented a model of collaborative planning in dialogue and discussed a number of parameters that can affect either the efficacy of the final plan or the efficiency of the collaborative planning process. Then in Section 5, I presented the results of experiments testing hypotheses about the effects of these parameters on collaborative planning dialogues. These results contribute to the development of the model of collaborative planning dialogue presented here. In addition, since the testbed implementation is compatible with many current theories, these results could be easily incorporated into other dialogue planning algorithms [17,48,51,53,79,88,127], *inter alia*.

A secondary goal of this paper was to argue for a particular methodology for dialogue theory development. The method was specified in Section 4.1. The Design-World testbed was introduced in Section 4 and Sections 4.4 and 4.5 described the parameterizations of the model that support testing the hypotheses. Four parameters for communicative strategies were tested: (1) All-implicit; (2) Close-consequence; (3) Explicit-warrant; and (4) Matched-pair-inference-explicit. Four parameters for tasks were tested: (1) Standard; (2) Zero-non-matching-beliefs; (3) Matched-pair (MP); (4) Zero-invalids. Three situations of varying processing effort were tested.

In this section, I will first summarize the hypotheses and the experimental results in Section 6.1, then I will discuss how the experimental results might generalize to situations not implemented in the testbed. Section 6.4 proposes future work and Section 6.5 consists of concluding remarks.

6.1. *Summary of results*

The hypotheses that were generated by the statistical analysis of the dialogue corpora are repeated below for convenience from Sections 2 and 4.6.

- HYPOTH-C1: agents produce Consequence IRUs to demonstrate that they made the inference that is made explicit.
- HYPOTH-C2: agents choose to produce Consequence IRUs to ensure that the other agent has access to inferrable information.
- HYPOTH-C3: the choice to produce a Consequence IRU is directly related to a measure of "how hard" the inference is.
- HYPOTH-C4: the choice to produce a Consequence IRU is directly related to a measure of "how important" the inference is.
- HYPOTH-C5: the choice to produce a Consequence IRU is directly related to the degree to which the task requires agents to be coordinated on the inferences that they have made.
- HYPOTH-A1: agents produce Attention IRUs to support the processes of deliberating beliefs and intentions.
- HYPOTH-A2: there is a DISCOURSE INFERENCE CONSTRAINT whose effect is that inferences in dialogue are derived from propositions that are currently discourse salient (in working memory).
- HYPOTH-A3: the choice to produce an Attention IRU is related to the degree of inferential complexity of a task as measured by the number of premises required to make task-related inferences.
- HYPOTH-A4: the choice to produce an Attention IRU is related to the degree to which an agent is resource limited in attentional capacity.
- HYPOTH-A5: the choice to produce an Attention IRU is related to the degree to which the task requires agents to be coordinated on the inferences that they have made.
- HYPOTH-I1: strategies that reduce collaborative effort without affecting quality of solution are beneficial.

Below I will summarize the experimental results reported in Section 5 with respect to the hypotheses above.

Hypotheses C3 and C4 were tested by comparing the Close-consequence strategy with the All-implicit strategy in the Standard task. In this experimental setup, the inference made explicit by the Consequence IRU was neither hard to make nor critical for performance. Hypothesis C3 was only weakly tested by the experiments because agents always make this inference. The results in Fig. 19 show that the Close-consequence strategy is detrimental for LOW AWM agents. This is because IRUs can displace useful information from working memory and because the inference made explicit with this IRU is not "hard enough".

The Standard task also provides a weak test of hypothesis C4. The fact that the Standard task is fault tolerant means that making the inference is not as critical as it might be. However, errors can result from either not making the inference or forgetting it once it is made. At lower values of AWM, the probability of such errors is not that high. However, the results shown in Fig. 19 show that the probability of error is higher for HIGH AWM agents in this case, because of their belief deliberation algorithm, and thus the Close-consequence strategy is beneficial for HIGH AWM agents, even in the Standard task.

The Zero-invalids task provides another test of hypothesis C4 by increasing the importance of the inference made explicit by the Close-consequence strategy. Fig. 25 shows that hypothesis C4 is confirmed because the Close-consequence strategy is beneficial for LOW, MID and HIGH AWM agents. In addition to the reasons discussed for the Standard task, this strategy is beneficial for HIGH AWM agents because they have more potential to improve their scores by ensuring that they don't make errors.

The experiments did not test hypothesis C1 because agents in the testbed are not designed to actively monitor evidence from other agents as to what inferences they might have made. Hypothesis C5 was not tested by the experiments because agents always rectify the situation if they detect a discrepancy in beliefs about act-effect inferences: they reject proposals whose preconditions do not hold.

Hypotheses A1, A4 and A5 were tested by experiments in which the Explicit-warrant strategy was compared with the All-implicit strategy in the Standard task. Hypothesis A1 is disconfirmed for LOW AWM agents. Fig. 16 shows that the Explicit-warrant strategy is neither beneficial nor detrimental for LOW AWM agents for the Standard task, when processing is free. This counter-intuitive result arises because, when agents are highly resource limited, IRUs can displace other information that is more useful.

To test hypothesis I1 in this situation, we also examined two situations where processing is not free. When communication cost dominates other processing costs, the Explicit-warrant strategy is detrimental for LOW and HIGH AWM agents. However, when retrieval cost dominates other processing costs, the Explicit-warrant strategy is beneficial for MID AWM agents and there is a trend toward a beneficial effect for HIGH AWM agents. Thus these two situations show that hypothesis I1 is confirmed: processing effort has a major effect on whether a strategy is beneficial.

We also tested hypotheses A1, A4 and A5 with experiments in which the Explicit-warrant strategy was compared with the All-implicit strategy in the Zero-non-matching-beliefs task (see Figs. 20 and 21). This task increases the importance of making deliberation-based inferences by requiring agents to be coordinated on these inferences in order to do well on the task. In this situation, we saw a very large beneficial effect for the Explicit-warrant strategy, which is not diminished by increasing communication effort. Thus in situations in which agents are required to be coordinated on these inferences, strategies which include Attention IRUs can be very important.

Hypotheses A2, A3, A4, and A5 were tested by experiments comparing the Matched-pair-inference-explicit strategy with the All-implicit strategy in the two versions of the Matched-pair task. The results shown in Figs. 22 and 23 provide support for these hypotheses. However these results also included an unpredicted benefit of Attention IRUs for inferentially complex tasks where agents must coordinate on inferences. Fig. 23 shows that both MID and HIGH AWM agents' performance improves with the Matched-pair-inference-explicit strategy. This can be explained by the fact that Attention IRUs increase the likelihood that agents will make the *same* inference, rather than *divergent* inferences, when multiple inferences are possible.

Furthermore, although the Matched-pair-inference-explicit strategy is specifically tied to Matched-pair inferences, it provides a test of a general strategy for making premises for inferences salient, when tasks are inferentially complex and require agents to remain coordinated on inferences. Thus it provides strong support for the DISCOURSE INFERENCE CONSTRAINT. To generalize this strategy to other cases of plan-related inferences, the clauses in the strategy plan operator that specifically refer to Matched-pair inferences can be replaced with a more general inference, e.g. the more general (Generates ?act1 $\wedge$?act2 ?act3), where the generates relation is to be inferred [33,51,97].

Hypothesis I1 was tested by examining extremes in cost ratios for retrieval effort and communication effort whenever a hypothesis about the beneficial effects of IRUs was confirmed. Fig. 18 shows that high communication effort can make the Explicit-warrant strategy detrimental in the Standard task. Fig. 21 shows that high communication effort does not eliminate the benefits of the Explicit-warrant strategy in the Zero-non-matching-beliefs task. Fig. 24 shows that high communication effort does not eliminate the benefits of the Matched-pair-inference-explicit strategy in the Matched-pair-two-room task. Thus the strategy of making premises for inferences salient is robust against extremes in processing effort.

6.2. *Generalizability of the results*

This section addresses concerns raised in [54] that simulation is "experimentation in the small". Hanks writes that [54, Section 5.1.5]:

> The ultimate value—arguably the *only* value—of experimentation is to constrain or otherwise inform the designer of a system that solves interesting problems. In order to do so the experimenter must demonstrate three things:
> (1) that her results—the relationships she demonstrates between agent characteristics and world characteristics—extend beyond the particular agent, world, and problem specification she studied,
> (2) that the solution to the problem area she studied in isolation will be applicable when that same problem area is encountered in a larger, more complex world, and
> (3) that the relationship demonstrated experimentally actually constrains or somehow guides the design of a larger more realistic agent.

The list in (1)-(3) are all different ways of saying that the results should generalize beyond the specifics of the experiment, and this after all is a basic issue with all experimental work. Typically generalizations can be shown by a series of multiple experiments modifying multiple variables as we have done here. For example, the modifications to the task are specifically designed to test whether beneficial communicative strategies generalize across tasks. However, we might also ask to what extent do the variables manipulated in the simulation abstract out key properties of real situations? Below I will briefly discuss why the results presented above are potentially generalizable. I will focus on generalizations along three dimensions: (1) task (or environmental) properties; (2) agent architectural properties; and (3) agent behaviors. These dimensions are the same as those in Cohen's "ecological triangle" [24].

Generalizations about tasks

The Design-World task was selected as a simple planning task that requires negotiation of each step. The structure of this task is isomorphic to a subcomponent of many collaborative planning tasks. In addition, to test generalizability of hypothesized benefits across tasks, we examined more complex variants of the task by manipulating three abstract features: (1) inferential complexity as measured by the number of premises required for making a task-related inference and (2) degree of belief coordination required on intentions, inferences and beliefs underlying a plan; and (3) the task determinacy and fault tolerance of the plan. These general features can certainly be applied to other tasks in other domains. In fact it is difficult to think of a task or domain in which these features could not be applied.

Generalizations about agent properties

Design-World agents are artificial agents that are designed to model the resource limited qualities of human agents. The planning and deliberation aspects of human processing are modeled with the IRMA architecture, and resource limits on these processes are modeled by extending the IRMA architecture with a model of attention/working memory (AWM) which has been shown to model a limited but critical set of properties of human processing. The way that agents process dialogue is tied to the agent architecture.

The experimental results will extend to dialogues between artificial agents to the extent that those agents exhibit similar cognitive properties. Here, we looked at a resource bound on access to memory as modeled by a size of memory subset limit, however size is directly correlated to *time* to access memory. Artificial agents are often time limited in rapidly changing worlds, so it seems quite plausible that artificial agents would benefit from similar communicative strategies. For example, I would predict that agents in the Phoenix simulation testbed would benefit from the strategies discussed here [24]. In other work artificial agents do "make inferences explicit" by communicating to other agents partial computations when the other agent might have been able to make these computations [24, 36, 130]. In addition, defining inferential complexity as a direct consequence of the number of premises simultaneously in memory bears a strong resemblance to problems artificial processors have when a computation requires a large working set [122].

The experimental results should extend to dialogues between humans and artificial agents because Design-World agents are designed to model humans. However it may be desirable to change the definition of collaborative effort for modeling human-computer interaction to allow the computer to handle processing that is easy for the computer to do and for the human to handle processing that is easy for the human to do. Furthermore, most of the claims about the AWM model are based on a limited set of human working memory properties, and these properties will also hold for other cognitively based architectures such as SOAR [74,77].

Generalizations about agent behaviors

In this work the agent behaviors that were tested were the agent communication strategies. One reason to believe that the strategies are general to human-human discourse is that they were based on observed strategies in different corpora of natural collaborative planning dialogues. It is possible to find all three types of IRUs in the Trains, Map-task and Design corpora [12,98,128,142], as well as in the financial advice domain.

In addition to this empirical evidence, there are further reasons why we might expect generalizations.

The communicative acts and discourse acts used by Design-World agents are similar to those used in [12,14,113,120]. Thus communicative strategies based on these acts should be implementable in any of these systems.

The experimental results based on these strategies should generalize to other discourse situations because the strategies are based on general relations between utterance acts and underlying processes, such as supporting deliberation and inference. For example, the mapping of a WARRANT relation between an act and a belief in naturally occurring examples such as (9) was modeled with a WARRANT relation between an act and a belief in Design-World, as seen in the Explicit-warrant communication strategy. The claims made about the use of the Explicit-warrant communication strategy should generalize to any dialogue planning domain where agents use warrants to support deliberation.

Similarly, content-based inferences in natural dialogues such as that discussed in relation to example (11) were modeled with content-based inferences in Design-World such as those required for doing well on the Matched-pair tasks. This inferential situation was designed to test the DISCOURSE INFERENCE CONSTRAINT, that inferences in dialogue are restricted to premises that are currently salient. Both experimental and corpus-based evidence was provided in support of the discourse inference constraint. The claims made about the use of the Matched-pair-inference-explicit communication strategy, based on experimental evidence, should generalize to any dialogue strategy where agents make premises for inferences available, and to any planning domain where agents are required to make content-based inferences in support of deliberation or planning.

The evaluation metrics applied to these strategies should also generalize whenever domain plan utility is a reasonable measure of the quality of solution for a dialogue task.

6.3. Relation to other work

The model of collaborative planning dialogues presented in Section 3 draws from previous work on cooperative dialogue [10,21,30,37,50,51,67,85,97,98,141,143], and the results are applicable to other current research on collaborative planning [17, 26, 32, 48, 53, 56, 87, 113, 127, 146].

The agent architecture and the model of deliberation and means–end reasoning is based on the work of [7, 34], and on Pollack's TileWorld simulation environment [99]. The use of IRMA as an underlying model of intention deliberation to provide a basis for a collaborative planning model was first proposed in [131–133], and has been incorporated into other work [48, 146]. The architecture includes a specific model of limited working memory, but most of the claims about the model are based on its recency and frequency properties, which might also be provided by other cognitively based architectures such as SOAR [74, 77].[22] Since the testbed architecture is consistent with that assumed in other work, the experimental results should be generalizable to those frameworks.

The relationship between discourse acts and domain-based options and intentions in this work is based on Litman's model of discourse plans [85, 86] and is similar to the approach in [12, 14, 127]. The emphasis on autonomy at each stage of the planning process and the belief reasoning mechanism of Design-World agents is based on the theory of belief revision and the multi-agent simulation environment developed in the Automated Librarian project [14, 38, 40, 88].

The Design-World testbed is based on the methods used in the TileWorld and Phoenix simulation environments: rapidly changing robot worlds in which an artificial agent attempts to optimize reasoning and planning [24, 54, 99]. TileWorld is a single agent world in which the agent interacts with its environment, rather than with another agent. Design-World uses similar methods to test a theory of the effect of resource limits on communicative behavior between two agents.

Design-World is also based on the method used in Carletta's JAM simulation for the Edinburgh Map-task [12, 100]. JAM is based on the Map-task dialogue corpus, where the goal of the task is for the planning agent, the instructor, to instruct the reactive agent, the instructee, how to get from one place to another on the map. JAM focuses on efficient strategies for recovery from error and parameterizes agents according to their communicative and error recovery strategies. Given good error recovery strategies, Carletta argues that "high risk" communicative strategies are more efficient, but did not attempt to quantify efficiency. In contrast, the approach here provides a way of quantifying what is an effective or efficient strategy, and the results suggest that a combination of the agents' resource limitations and the task definition determine when strategies are efficient. Future work could test Carletta's claims about recovery strategies within this extended framework.

To my knowledge, none of this earlier work has considered the factors that affect the range of variation in communicative choice, or the effects of different choices, or measured how communicative choice affects the construction of a collaborative plan and the ability of the conversants to stay coordinated. Nor have other theories of collaborative planning been explicit about the agent architecture, or tested specific ideas about

[22] [136] discusses the differences between an AWM-like attentional model and Grosz and Sidner's stack model of attentional state [49, 50, 114]. See also [107] for a discussion of other discourse phenomena for which the AWM model may be useful.

resource bounds in dialogue, and none have used utility as the basis for agents' communicative choice. In addition, no earlier work on cooperative task-oriented dialogue argued that conversational agents' resource limits and task complexity are major factors in determining effective conversational strategies in collaboration.

6.4. Future work

A promising avenue for future work is to investigate beneficial strategies for teams of heterogeneous agents. In the experiments here, pairs of agents in dialogue were always parameterized with the same resource limits. Pilot studies of dialogues between heterogeneous agents suggest that strategies that are not effective for homogeneous agents may be effective for heterogeneous ones. For example, in [133] I tested an Attention IRU strategy in which agents would tell one another about all the options they knew about at the beginning of planning each room. This strategy is not beneficial for homogeneous agents because IRUs can displace other useful information. However if one agent is not limited, then it can be helpful for the resource limited agent to exploit the capabilities of the more capable agent by telling the other agent important facts before it forgets them.

Another extension would be to extend the agent communication strategies or to test additional ones. For example, other work proposes a number of strategies for information selection and ordering in dialogue and provides some evidence that these strategies are efficient or efficacious [11, 17, 123, 147]. Support for these claims could be provided by Design-World experiments in which agents used these strategies to communicate.

Future work could also modify the properties of the world or of the task. For example, it would be possible to make Design-World more like TileWorld by making the world change in the course of the task, by adding or removing furniture.

These results may also be incorporated as input into decision algorithms in which agents decide online which strategy to pursue, and investigate additional factors that determine when strategies are effective in collaborative planning dialogues. The results presented here show what information an agent should consider. For example, a comparison between LOW, MID and HIGH AWM agents shows how to design decision algorithms for agents who have to decide whether to expend additional effort.

Another promising avenue is make the agents capable of remembering and learning from past mistakes so that they can adapt their strategies to the situation [3].

Finally, these results should be incorporated into the design of multi-agent problem-solving systems and into systems for human–computer communication, such as those for teaching, advice and explanation, where for example the use of particular strategies might be premised on the abilities of the learner or apprentice.

6.5. Concluding remarks

The goal of this paper was to show how agents' choice in communicative action, their algorithms for language behavior, can be designed to mitigate the effect of their resource limits in the context of particular features of a collaborative planning task. In this paper, I first motivate a number of hypotheses based on a statistical analysis of natural collaborative planning dialogues. Then a functional model of collaborative planning dialogues is developed based on these hypotheses, including parameters that are hypothesized to affect the generalizability of the model. The model is then implemented in a testbed in which these parameters can be varied, and the hypotheses are tested.

The method used here can be contrasted with other work on dialogue modeling. Much previous work on dialogue modeling only carries out part of the process described above: only the initial part of the process up to specifying a functional model is completed. Follow-on research that is based on these models must judge the model according to subjective criteria such as how well it fits researcher's intuitions or how elegant the model is. The models developed here on the basis of empirical evidence can also be judged according to these subjective criteria, but this work carries out additional steps to further test and refine the model suggested by the corpus analysis. Implementing a model with parameters to test the generalizability of the model and testing hypotheses in a testbed implementation provides a way to check subjective evaluations and suggests many ways in which our initial hypotheses must be refined and further tested.

The Design-World testbed is the first testbed for conversational systems that systematically introduces several different types of independent parameters that are hypothesized to affect the efficacy of a collaborative plan negotiated through a dialogue, and the efficiency of that dialogue process. Experiments in the testbed examined the interaction between (1) agents' resource limits in attentional capacity and inferential capacity; (2) agents' choice in communication; and (3) features of communicative tasks that affect task difficulty such as inferential complexity, degree of belief coordination required, and tolerance for errors. The results verified a number of hypotheses that depended on particular assumptions about agents' resource limits that were not possible to test by corpus analysis alone.

Several unpredicted and counter-intuitive results were also demonstrated by the experiments. First, the task property of belief coordination in *combination* with resource limits (as in the Zero-non-matching-beliefs and Matched-pair tasks), were shown to produce the most robust benefits for IRUs, rather than resource limits alone as originally hypothesized. Second, I predicted that IRUs would always be beneficial for LOW AWM agents, but found that IRUs can be detrimental for these agents through a side effect of displacing other, more useful, beliefs from working memory. Third, it would seem plausible that HIGH AWM agents should always perform better than either LOW or MID AWM agents since these agents always have access to more information. However the results showed that there are two situations in which this is not an advantage: (1) when accessing information has some cost; and (2) when access to multiple beliefs can lead agents to make divergent inferences. In this case, restricting agents to a small shared working set is a natural way to limit inferential processes. This limit intuitively corresponds to potential benefits of limited working memory for humans and explains how humans manage to coordinate on inferences in conversation [49, 66, 83].

These results clearly demonstrate that factors not previously considered in dialogue models must be taken into account of claims if cooperativity, efficiency, or efficacy are to be supported. In addition, I have shown that a theory of dialogue that includes a model of resource limited processing can account for both the observed language behavior in human–human dialogue and the experimental results presented here.

Acknowledgements

The work reported in this paper has benefited from discussions with Steve Whittaker, Aravind Joshi, Ellen Prince, Mark Liberman, Max Mintz, Bonnie Webber, Scott Weinstein, Candy Sidner, Owen Rambow, Beth Ann Hockey, Karen Sparck Jones, Julia Galliers, Phil Stenton, Megan Moser, Johanna Moore, Christine Nakatani, Penni Sibun, Ellen Germain, Janet Cahn, Jean Carletta, Jon Oberlander, Julia Hirschberg, Alison Cawsey, Rich Thomason, Cynthia McLemore, Jerry Hobbs, Pam Jordan, Barbara Di Eugenio, Susan Brennan, Rebecca Passonneau, Rick Alterman and Paul Cohen. I am grateful to Julia Galliers for providing me with an early implementation of the belief revision mechanism used in the Automated Librarian project, and to Julia Hirschberg who provided me with tapes of the financial advice talk show. Thanks also to the two anonymous reviewers who provided many useful suggestions.

This research was partially funded by ARO grants DAAG29-84-K-0061 and DAAL03-89-C0031PRI, DARPA grants N00014-85-K0018 and N00014-90-J-1863, NSF grants MCS-82-19196 and IRI 90-16592 and Fellowship INT-9110856 for the 1991 Summer Science and Engineering Institute in Japan, and Ben Franklin Grant 91S.3078C-1 at the University of Pennsylvania, and by Hewlett-Packard Laboratories.

References

[1] J.F. Allen, Recognizing intentions from natural language utterances, in: M. Brady and R.C. Berwick, eds., *Computational Models of Discourse* (MIT Press, Cambridge, MA, 1983).

[2] J.F. Allen and C.R. Perrault, Analyzing intention in utterances, *Artif. Intell.* **15** (1988) 143–178.

[3] R. Alterman, R. Zito-Wolf and T. Carpenter, Interaction, comprehension and instruction usage, *J. Learn. Sci.* **1** (1991) 273–318.

[4] J.R. Anderson and G.H. Bower, *Human Associative Memory* (Winston and Sons, 1973).

[5] A. Baddeley, *Working Memory* (Oxford University Press, Oxford, 1986).

[6] J. Barwise, The situation in logic—IV: on the model theory of common knowledge, Tech. Rept. No. 122, Center for the Study of Language and Information, Ventura Hall, Stanford University, Stanford, CA (1988).

[7] M. Bratman, D. Israel and M. Pollack, Plans and resource bounded practical reasoning, *Comput. Intell.* **4** (1988) 349–355.

[8] S.E. Brennan, Seeking and providing evidence for mutual understanding, Ph.D. Thesis, Psychology Department, Stanford University, Stanford, CA (1990).

[9] S.E. Brennan and E.A. Hulteen, Interaction and feedback in a spoken language system, *Knowledge Based Systems* (1995).

[10] S. Carberry, Plan recognition and its use in understanding dialogue, in: A. Kobsa and W. Wahlster, eds., *User Models in Dialogue Systems* (Springer, Berlin, 1989) 133–162.

[11] G. Carenini and J. Moore, Generating explanation in context, in: *Proceedings International Workshop on Intelligent User Interfaces* (1993).

[12] J.C. Carletta, Risk taking and recovery in task-oriented dialogue, Ph.D. Thesis, Edinburgh University, Edinburgh (1992).

[13] A. Cawsey, Generating interactive explanations, in: *Proceedings AAAI-91*, Anaheim, CA (1991) 86–91.

[14] A. Cawsey, J. Galliers, S. Reece and K. Sparck Jones, Automating the librarian: a fundamental approach using belief revision, Tech. Rept. 243, Cambridge Computer Laboratory, Cambridge, MA (1992).

[15] A. Chapanis, Interactive human communication, *Sci. Amer.* **232** (1975) 34–42.

[16] J. Chu-Carrol and S. Carberry, A plan-based model for response generation in collaborative task-oriented dialogue, in: *Proceedings AAAI-94*, Seattle, WA (1994) 799–805.

[17] J. Chu-Carrol and S. Carberry, Response generation in collaborative negotiation, in: *Proceedings 33rd Annual Meeting of the Association for Computational Linguistics* (1995).

[18] H.H. Clark and S.E. Brennan, Grounding in communication, in: L.B. Resnick, J. Levine and S.D. Bahrend, eds., *Perspectives on Socially Shared Cognition* (APA, 1990).

[19] H.H. Clark and C.R. Marshall, Definite reference and mutual knowledge, in: A.K Joshi, B.L. Webber and I. Sag, eds., *Elements of Discourse Understanding* (Cambridge University Press, Cambridge, 1981) 10-63.

[20] H.H. Clark and E.F. Schaefer, Collaborating on contributions to conversations, *Lang. Cognit. Processes* **2** (1987) 19-41.

[21] H.H. Clark and E.F. Schaefer, Contributing to discourse, *Cognit. Sci.* **13** (1989) 259-294.

[22] H.H. Clark and D. Wilkes-Gibbs, Referring as a collaborative process, *Cognition* **22** (1986) 1-39.

[23] P.R. Cohen, A survey of the Eighth National Conference on Artifical Intelligence: pulling together or pulling apart, *AI Magazine* **12** (1) (1991) 16-41.

[24] P.R. Cohen, M.I. Greenburg, D.M. Hart and A.E. Howe, Trial by fire: requirements for agents in complex environments, *AI Magazine* **10** (3) (1989) 32-48.

[25] P.R. Cohen, *Empirical Methods for Artificial Intelligence* (MIT Press, Boston, MA, 1995).

[26] P. Cohen and H. Levesque, Teamwork, *Nous* **25** (1991) 11-24.

[27] P.R. Cohen, On knowing what to say: planning speech acts, Tech. Rept. 118, Department of Computer Science, University of Toronto, Toronto, Ont. (1978).

[28] P.R. Cohen, The pragmatics of referring and the modality of communication, *Comput. Linguist.* **10** (1984) 97-146.

[29] P.R. Cohen, *Reasoning about Uncertainty: An Artificial Intelligence Approach* (Pitman, Boston, MA, 1985).

[30] R. Cohen, Analyzing the structure of argumentative discourse, *Comput. Linguist.* **13** (1987) 11-24.

[31] A.M. Collins and M.R. Quillian, Retrieval time from semantic memory, *J. Verbal Learn. Verbal Behav.* **8** (1969) 240-247.

[32] N. Dahlback, Representations of discourse: cognitive and computational aspects, Ph.D. Thesis, Linköping University, Linköping (1991).

[33] B. Di Eugenio, Understanding natural language instructions: a computational approach to purpose clauses, Ph.D. Thesis, University of Pennsylvania, Philadelphia, PA (1993).

[34] J. Doyle, Rationality and its roles in reasoning, *Comput. Intell.* **18** (3) (1992).

[35] E.H. Durfee, P. Gmytrasiewics and J. Rosenschein, The utility of embedded communications and the emergence of protocols, in: *Proceedings AAAI Workshop on Planning for Interagent Communication* (1994).

[36] E.H. Durfee, V.R. Lesser and D.D. Corkill, Cooperation through communication in a distributed problem solving network, in: *Distributed Artificial Intelligence* (Morgan Kaufmann, Los Altos, CA, 1987).

[37] T.W. Finin, A.K. Joshi and B.L. Webber, Natural language interactions with artificial experts, *Proc. IEEE* **74** (1986) 921-938.

[38] J.R. Galliers, A theoretical framework for computer models of cooperative dialogue, acknowledging multi agent conflict, Tech. Rept. 172, Computer Laboratory, University of Cambridge, Cambridge (1989).

[39] J.R. Galliers, Belief revision and a theory of communication, Tech. Rept. 193, Computer Laboratory, University of Cambridge, Cambridge (1990).

[40] J.R. Galliers, Autonomous belief revision and communication, in: P. Gärdenfors, ed., *Belief Revision* (Cambridge University Press, Cambridge, 1991) 220-246.

[41] J.R. Galliers, Cooperative interaction as strategic belief revision, in: M.S. Deen, ed., *Cooperating Knowledge Based Systems* (Springer, Berlin, 1991) 148-163.

[42] P. Gärdenfors, *Knowledge in Flux : Modeling the Dynamics of Epistemic States* (MIT Press, Cambridge, MA, 1988).

[43] P. Gardenfors, The dynamics of belief systems: foundations vs. coherence theories, *Revue International De Philosophie* (1990).

[44] G. Gazdar, *Pragmatics: Implicature, Presupposition and Logical Form* (Academic Press, New York, 1979).

[45] A.I. Goldman, *Epistemology and Cognition* (Harvard University Press, Cambridge, MA, 1986).

[46] H.P. Grice, *William James Lectures* (1967).
[47] H.P. Grice, Logic and conversation, in: P. Cole and J. Morgan, eds., *Syntax and Semantics III—Speech Acts* (Academic Press, New York, 1975) 41-58.
[48] B. Grosz and S. Kraus, Collaborative plans for group activities, in: *Proceedings IJCAI-93*, Chambery (1993).
[49] B.J. Grosz, The representation and use of focus in dialogue understanding, Tech. Rept. 151, SRI International, Menlo Park, CA (1977).
[50] B.J. Grosz and C.L. Sidner, Attentions, intentions and the structure of discourse, *Comput. Linguist.* **12** (1986) 175-204.
[51] B.J. Grosz and C.L. Sidner, Plans for discourse, in: P.R. Cohen, J. Morgan and M. Pollack, eds., *Intentions in Communication* (MIT Press, Cambridge, MA, 1990).
[52] C.I. Guinn, A computational model of dialogue initiative in collaborative discourse, AAI Tech. Rept. FS-93-05 (1993).
[53] C.I. Guinn, Meta-dialogue behaviors: improving the efficiency of human-machine dialogue, Ph.D. Thesis, Duke University, Durham, NC (1994).
[54] S. Hanks, M. Pollack and P. Cohen, Benchmarks, testbeds, controlled experimentation and the design of agent architectures, *AI Magazine* **14** (1993)
[55] B. Hayes-Roth and P.W. Thorndyke, Integration of knowledge from text, *J. Verbal Learn. Verbal Behav.* **18** (1979) 91-108.
[56] P.A. Heeman and G. Hirst, Collaborating on referring expressions, *Comput. Linguist.* **21** (1995) 351-383.
[57] S. Hellyer, Frequency of stimulus presentation and short-term decrement in recall, *J. Exper. Psych.* **64** (1962) 650.
[58] D.L. Hintzmann and R.A. Block, Repetition and memory: evidence for a multiple trace hypothesis, *J. Exper. Psych.* **88** (1971) 297-306.
[59] J. Hirschberg, A theory of scalar implicature, Ph.D. Thesis, Computer and Information Science, University of Pennsylvania, Philadelphia, PA (1985).
[60] J.R. Hobbs, On the coherence and structure of discourse, Tech. Rept. CSLI-85-37, Center for the Study of Language and Information, Ventura Hall, Stanford University, Stanford, CA (1985).
[61] J.R. Hobbs, Intention, information and structure in discourse, in: D. Scott and E. Hovy, eds., *Burning Issues in Discourse* (1994).
[62] A. Jameson, Knowing what others know, Ph.D. Thesis, University of Nijmegen, Nijmegen (1990).
[63] P.N. Johnson-Laird, *Deduction* (Lawrence Erlbaum, Hillsdale, NJ, 1991).
[64] P. Jordan and M.A. Walker, Multi-agent resource sensitive communication, in: *Proceedings AAAI Fall Workshop on Rational Agency* (1995)
[65] A.K. Joshi, Informationally equivalent sentences, in: *Proceedings International Conference on Microwaves, Circuit Theory and Information Theory* (1964).
[66] A.K. Joshi, Some extensions of a system for inference on partial information, in: *Pattern Directed Inference Systems* (Academic Press, New York, 1978) 241-257.
[67] A.K. Joshi, B. Webber and R.M. Weischedel, Some aspects of default reasoning in interactive discourse, Tech. Rept. MS-CIS-86-27, University of Pennsylvania, Philadelphia, PA (1986).
[68] A.K. Joshi, B.L. Webber and R.M. Weischedel, Preventing false inferences, in: *Proceedings Tenth International Conference on Computational Linguistics*, Palo Alto, CA (1984) 134-138.
[69] L. Karttunen and S. Peters, Conventional implicature, *Syntax Semant.* **11** (1979).
[70] G. Keppel, *Design and Analysis: A Researcher's Handbook* (Prentice-Hall, Englewood Cliffs, NJ, 2nd ed., 1982).
[71] A. Kidd, The consultative role of an expert system, in: P. Johnson and S. Cook, eds., *People and Computers: Designing the Interface* (Cambridge University Press, Cambridge, 1985).
[72] K. Konolige, Belief and incompleteness, in: J.R. Hobbs and R.C. Moore, eds., *Formal Theories of the Commonsense World* (Ablex, Norwood, NJ, 1985) 359-403.
[73] K. Konolige, *A Deduction Model of Belief* (Brown University Press, Providence, RI, 1986).
[74] J.E. Laird, A. Newell and P.S. Rosenbloom, SOAR: an architecture for general intelligence, *Artif. Intell.* **33** (1987) 1-64.
[75] T.K. Landauer, Reinforcement as consolidation, *Psych. Rev.* **16** (1969) 82-96.

[76] T.K. Landauer, Memory without organization: properties of a model with random storage and undirected retrieval, *Cognit. Psych.* **7** (1975) 495–531.

[77] J.F. Lehman, R.L. Lewis and A. Newell, Natural language comprehension in soar, Tech. Rept. CMU-CS-91-117, Carnegie Mellon University, Pittsburgh, PA (1991).

[78] N. Lenke, Anticipating the reader's problmes and the automatic generation of paraphrases, in: *Proceedings COLING-94*, Kyoto (1994).

[79] H.J. Levesque, P.R. Cohen and J.H.T. Nunes, On acting together, in: *Proceedings AAAI-90*, Boston, MA (1990).

[80] S.C. Levinson, Activity types and language, *Linguistics* **17** (1979) 365–399.

[81] S.C. Levinson, Some pre-observations on the modelling of dialogue, *Discourse Processes* **4** (1981) 93–116.

[82] S.C. Levinson, *Pragmatics* (Cambridge University Press, Cambridge, 1983).

[83] S.C. Levinson, What's special about conversational inference, *1987 Linguistics Institute Packet* (1985).

[84] D. Lewis, *Convention* (Harvard University Press, Cambridge, MA, 1969).

[85] D. Litman, Plan recognition and discourse analysis: an integrated approach for understanding dialogues, Tech. Rept. 170, University of Rochester, Rochester, NY (1985).

[86] D. Litman and J. Allen, Recognizing and relating discourse intentions and task-oriented plans, in: P.R. Cohen, J. Morgan and M. Pollack, eds., *Intentions in Communication* (MIT Press, Cambridge, MA, 1990).

[87] K.E. Lochbaum, Using collaborative plans to model the intentional structure of discourse, Ph.D. Thesis, Harvard University, Cambridge, MA (1994).

[88] B. Logan, S. Reece and K. Sparck Jones, Modelling information retrieval agents with belief revision, in: *Proceedings Seventh Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Springer, London, 1994) 91–100.

[89] W.C. Mann and S.A. Thompson, Rhetorical structure theory: description and construction of text structures, in: G. Kempen, ed., *Natural Language Generation* (Martinus Nijhoff, Dordrecht, 1987) 83–96.

[90] K.R. McKeown, Paraphrasing questions using given and new information, *Comput. Linguist.* **9** (1983).

[91] G.A. Miller, The magical number seven, plus or minus two: some limits on our capacity for processing information, *Psych. Rev.* **3** (1956) 81–97.

[92] J.D. Moore and C.L. Paris, Planning text for advisory dialogues: capturing intentional and rhetorical information, *Comput. Linguist.* **19** (4) (1993).

[93] M.G. Moser and J. Moore, in: *Proceedings ACL Workshop on Intentionality and Structure in Discourse Relations* (1993).

[94] M.G. Moser and J. Moore, Investigating cue selection and placement in tutorial discourse, in: *Proceedings 33rd Annual Meeting of the Association for Computational Linguistics* (1995).

[95] D.A. Norman and D.G. Bobrow, On data-limited and resource-limited processes, *Cognit. Psych.* **7** (1975) 44–46.

[96] R. Perrault, An application of default logic to speech-act theory, in: P.R. Cohen, J. Morgan and M. Pollack, eds., *Intentions in Communication* (MIT Press, Cambridge, MA, 1990).

[97] M. Pollack, Inferring domain plans in question answering, Tech. Rept. 403, SRI International - Artificial Intelligence Center, University of Pennsylvania, Philadelphia, PA (1986).

[98] M. Pollack, J. Hirschberg and B. Webber, User participation in the reasoning process of expert systems, in: *Proceedings AAAI-82*, Pittsburgh, PA (1982).

[99] M. Pollack and M. Ringuette, Introducing the TileWorld: experimentally evaluating agent architectures, in: *Proceedings AAAI-90*, Boston, MA (1990) 183–189.

[100] R. Power, A computer model of conversation, Ph.D. Thesis, University of Edinburgh, Edinburgh (1974).

[101] R. Power, Mutual intention, *J. Theory Social Behav.* **14** (1984).

[102] E.F. Prince, Toward a taxonomy of given-new information, in: *Radical Pragmatics* (Academic Press, New York, 1981) 223–255.

[103] E.F. Prince, The ZPG letter: subjects, definiteness and information status, in: S. Thompson and W. Mann, eds., *Discourse Description: Diverse Analyses of a Fund Raising Text* (Benjamin, New York, 1992) 295–325.

[104] R. Reichman, *Getting Computers to Talk Like You and Me* (MIT Press, Cambridge, MA, 1985).

[105] R. Reiter, A logic for default reasoning, *Artif. Intell.* **13** (1980) 81-132.
[106] N. Reithinger and E. Maier, Utilizing statistical speech act processing in verbmobil, in: *Proceedings 33rd Annual Meeting of the Association for Computational Linguistics* (1995).
[107] C.P. Rose, B. Di Eugenio, L. Levin and C. Van Ess-Dykema, Discourse processing of dialogues with multiple threads, in: *Proceedings 33rd Annual Meeting of the Association for Computational Linguistics* (1995).
[108] J.S. Rosenschein, Rational interaction and multi agent planning, Ph.D. Thesis, Computer Science Department, Stanford University, Stanford, CA (1985)
[109] J.M. Sadock, On testing for conversational implicature, in: P. Cole, ed., *Syntax and Semantics* **9**: *Pragmatics* (Academic Press, New York, 1978) 281-297.
[110] E.A. Schegloff, Between micro and macro: contexts and other connections, in: *The Micro-Macro Link* (University of California Press, Berkeley, CA, 1987).
[111] P. Sibun, The local organisation and incremental generation of text, Ph.D. Thesis, University of Massachusetts, Amherst, MA (1991)
[112] C. Sidner, Plan parsing for intended response recognition in discourse, *Comput. Intell.* **1** (1985).
[113] C. Sidner, An artificial discourse language for collaborative negotiation, in: *Proceedings AAAI-94*, Seattle, WA (1994) 814-820.
[114] C.L. Sidner, Toward a computational theory of definite anaphora comprehension in English, Tech. Rept. AI-TR-537, MIT, Cambridge, MA (1979).
[115] S. Siegel, *Nonparametric Statistics for the Behavioral Sciences* (McGraw-Hill, New York, 1956).
[116] R. Smith, The contract net protocol: high level communication and control in a distributed problem solver, *IEEE Trans. Comput.* **12** (1980) 1104-1113.
[117] R. Smith, D.R. Hipp and A.W. Biermann, A dialogue control algorithm and its performance, in: *Proceedings Third Conference on Natural Language Processing*, Trento (1992).
[118] M. Solomon, Scientific rationality and human reasoning, *Philos. Sci.* **59** (3) (1992).
[119] R.C. Stalnaker, Assertion, in: P. Cole, ed., *Syntax and Semantics* **9**: *Pragmatics* (Academic Press, New York, 1978) 315-332.
[120] A. Stein and U. Thiel, A conversational model of multimodal interaction, in: *Proceedings AAAI-93*, Washington, DC (1993).
[121] S. Sternberg, High-speed scanning in human memory, *Science* **153** (1967) 652-654.
[122] H.S. Stone, *High Performance Computer Architecture* (Addison-Wesley, Reading, MA, 1987).
[123] D.D. Suthers, Sequencing explanations to enhance communicative functionality, in: *Proceedings 15th Annual Conference of the Cognitive Science Society*, Boulder, CO (1993).
[124] R. Thomason, Accommodation, meaning and implicature: interdisciplinary foundations for pragmatics, in: P.R. Cohen, J. Morgan and M. Pollack, eds., *Intentions in Communication* (MIT Press, Cambridge, MA, 1990).
[125] R. Thomason, Propagating epistemic coordination through mutual defaults—I, in: R. Parikh, ed., *Proceedings Third Conference on Theoretical Aspects of Reasoning about Knowledge*, Pacific Grove, CA (1990) 29-39.
[126] D. Traum, Mental state in the trains-92 dialogue manager, in: *Proceedings AAAI Spring Symposium on Reasoning about Mental States* (1993).
[127] D. Traum, A computational model of grounding in natural language conversation, Ph.D. Thesis, University of Rochester, Rochester, NY (1994).
[128] D.R. Traum, The discourse reasoner in trains-90, TRAINS Technical Note 91-5, Department of Computer Science, University of Rochester, Rochester, NY (1991).
[129] E. Tulving, The effects of presentation and recall of material in free recall learning, *J. Verbal Learn. Verbal Behav.* **6** (1967).
[130] E. Turner, Selecting information to communicate, in: *Proceedings AAAI Workshop on Planning for Interagent Communication* (1994).
[131] M.A. Walker, Redundancy in collaborative dialogue, in: *Proceedings Fourteenth International Conference on Computational Linguistics* (1992) 345-351.
[132] M.A. Walker, Information and deliberation in discourse, in: *Proceedings ACL Workshop on Intentionality and Structure in Discourse Relations* (1993).

[133] M.A. Walker, Informational redundancy and resource bounds in dialogue, Ph.D. Thesis, University of Pennsylvania, Philadelphia, PA (1993).

[134] M.A. Walker, Discourse and deliberation: testing a collaborative strategy, in: *Proceedings COLING-94*, Kyoto (1994).

[135] M.A. Walker, Rejection by implicature, in: *Proceedings 20th Meeting of the Berkeley Linguistics Society* (1994).

[136] M.A. Walker, Limited attention and discourse structure, *Comput. Linguist.* (to appear).

[137] M.A. Walker, Testing collaborative strategies by computational simulation: cognitive and task effects, in: *Knowledge Based Systems* (1995).

[138] M.A. Walker and O. Rambow, The role of cognitive modeling in achieving communicative intentions, in: *Proceedings 7th International Conference on Natural Language Generation* (1994).

[139] M.A. Walker and S. Whittaker, Mixed initiative in dialogue: an investigation into discourse segmentation, in: *Proceedings 28th Annual Conference of the Association for Computational Linguistics*, Pittsburgh, PA (1990) 70-79.

[140] B.L. Webber, A formal approach to discourse anaphora, Ph.D. Thesis, Harvard University, Cambridge, MA (1978).

[141] B.L. Webber and A.K. Joshi, Taking the initiative in natural language database interaction: justifying why, in: *Proceedings COLING-82*, Prague (1982) 413-419.

[142] S. Whittaker, E. Geelhoed and E. Robinson, Shared workspaces: how do they work and when are they useful?, *Int. J. Man-Mach. Stud.* **39** (1993) 813-842.

[143] S. Whittaker and P. Stenton, Cues and control in expert client dialogues, in: *Proceedings 26th Annual Meeting of the Association of Computational Linguistics*, Buffalo, NY (1988) 123-130.

[144] S. Wilks, *Mathematical Statistics* (Wiley, New York, 1962).

[145] M.R. Young and J.D. Moore, Representation and generation in discourse planning, Tech. Report, University of Pittsburgh, Pittsburgh, PA (1994).

[146] M.R. Young, J.D. Moore and M.E. Pollack, Towards a principled representation of discourse plans, in: *Proceedings 16th Annual Conference of the Cognitive Science Society* (1994).

[147] I. Zukerman and R. McConachy, Generating concise discourses that addresses a user's inferences, in: *Proceedings IJCAI-93*, Chambery, France (1993).

[148] I. Zukerman and J. Pearl, Comprehension-driven generation of meta-technical utterances in math tutoring, in: *Proceedings AAAI-86*, Philadelphia, PA (1986) 606-611.