1

Contextual Influences on Attribute Selection for Repeated Descriptions

Pamela W. Jordan

Can a description that re-evokes a discourse entity satisfy multiple goals and if so what goals in addition to identification might there be? We propose and test five general goal types that reflect the functions of redundancy at the utterance level in a two part corpus analysis. We examine correlations between descriptions and contextual features indicative of the proposed goals and compare the performance of computer selection strategies to that of humans. Four out of the five general goal types we tested had an influence on the attributes included in redescriptions for the corpus we studied. We discuss how these results might apply to other types of task-oriented dialogues.

1.1 Introduction

In an extended discourse, speakers often redescribe objects that were introduced earlier in order to say something more about the object or the event in which it participates. As an object is described, the hearer and speaker create a discourse entity to relate the information about the object in the utterance to the appropriate mental representation of the object (Karttunen 1976, Webber 1978, Heim 1983, Kamp 1993, Passonneau 1996). The main goal when redescribing an entity is generating an expression that will efficiently and effectively re-evoke the appropriate discourse entity.

However, a goal-directed view of sentence generation suggests that speakers can attempt to satisfy multiple goals with each utterance (Appelt 1985) and that a single linguistic form can opportunistically con-

Information Sharing.
Kees van Deemter and Rodger Kibble (eds.).

tribute to the satisfaction of multiple goals (Stone and Webber 1998). The possibilities that goals besides identification could influence the content of a nominal expression and that an identification goal could be satisfied by more than a nominal expression have not yet been addressed in computational work on generating discourse anaphoric expressions.

The many-one mapping of goals to linguistic forms is more generally referred to as *overloading intentions* (Pollack 1991). Overloading can involve tradeoff across linguistic levels. For example, an intention which is achieved by complicating a form at the semantic level may allow the speaker to simplify at the syntactic level by omitting important information (Stone and Webber 1998).

Although we have learned that overloading is natural and perhaps even necessary, we have no well supported account of what degree of overloading is reasonable and what forms can more readily address multiple goals in dialogue. Without such an account, we have no principled way to deploy overloading in the automatic generation of natural language. Without well supported constraints on overloading, we are liable to create overloads in unnatural ways which will actually impede effective communication. For instance, we may produce descriptions and utterances that are too densely packed to be readily comprehensible by the hearer.

To begin an exploration of overloading, we analyzed a corpus of computer-mediated design dialogues which contain a large proportion of redescriptions that appear to be overspecified (Passonneau 1996, Vonk et al. 1992, O'Donnell et al. 1998). We define a *redescription*, in this work, as anything syntactically realized within an utterance that is mutually known and could be used to re-evoke a discourse entity. This includes nominal and pronominal expressions as well as adjectives within copulas, but excludes any information about a discourse entity that is new to the hearer (e.g. my desk chair is maple, it is maple). Overspecified redescriptions are those that provide more information about the discourse entity than is needed for identification purposes.

First we will describe the corpus we used in our analysis and then show that there are a large proportion of overspecified redescriptions in this corpus using a estimation procedure similar to that of (Passonneau 1996). Having a large proportion of overspecified redescriptions ensures that something besides identification is influencing the redescriptions in the corpus. We expect that the influences on redescriptions will vary with the type of task that is the topic of the dialogue and the communications setting and that there could be tasks and settings in which identification is the only prevalent influence on the descriptions (e.g. the tangram task (Clark and Schaefer 1989, Brennan 1990)).

Next we will describe five communication and problem solving inferences that could influence the choice of attributes used in redescriptions. We will describe their theoretical motivations and give examples from the corpus. In general they are motivated by the functions of repetition at the utterance or propositional level (Walker 1993, Johnstone 1994) and the inferences and implicit knowledge that bind natural language utterances together to form a coherent discourse (Grice 1975). Our main hypothesis is that attribute selections will be influenced by the contexts or situations in which we can expect inferences about joint commitments, changes to problem solving constraints, motivations for a proposal, closing a subtask and achieving understanding of task entity descriptions. The low-level definitions of the contexts in which we expect these inferences to occur are particular to the task addressed in the dialogues and the particular inference involved and are derived from sets of features annotated in the corpus. We will describe the features and how the contexts are recognized from the feature sets. We expect the types of inference we are considering to extend to other corpora, but that the applicability of each inference type will depend on the underlying task and communications setting of a corpus.

Finally, we will describe a two-part analysis of the corpus that explores our hypotheses about influences on redescriptions. The first part of the analysis examines the correlations between utterance and dialogue features that are indicative of when particular inferences are expected and the attributes expressed in redescriptions. The second part of the analysis utilizes computer simulations of attribute selection strategies. The data input to the selection strategies are the discourse entities evoked in the corpus and the contexts in which they occur. We measure the performance of a selection strategy by comparing the attributes it selects to those expressed in the corpus. We then compare the performances of two attribute selection strategies that consider only the identification goal with that of a selection strategy that considers both the identification goal and the inference contexts we described.

Within each selection strategy, parameters direct how supporting calculations for each strategy are made (e.g. how a distractor or context set is determined). Although these low-level adjustments in the strategies partially fit the selections to the corpus, our goal in this analysis is not to produce a general attribute selection algorithm for the domain of our corpus or any other corpus but to extract the best possible performance from each strategy before comparing them to one another. Our goal is to find out whether overloading applies to redescriptions and if so what general types of inferences or other communicative goals could potentially be overloaded in redescriptions.

The results of the two part analysis indicate that the first four of the five inference contexts we listed above have merit as possible influences on attribute selection for redescriptions. Neither part of the analysis indicated that an inference involving understanding would influence the content of a redescription in our corpus.

1.2 The COCONUT Corpus

Our analysis is based on the COCONUT corpus (Di Eugenio et al. 2000). This corpus contains 24 computer-mediated dialogues and we used 13 of these dialogues to test our hypotheses.[1] On average each dialogue in this subset of 13 contained 42 utterances, 25 discourse entities, 12 redescriptions and 6 utterances between redescriptions.

In each dialogue, two people collaborate on a simple design task; buying furniture for two rooms of a house. The information needed to complete the design task is divided between the two designers in such a way that a good design cannot be achieved without collaboration. With this task, the designers typically describe the furniture items that they believe are relevant to the current subtask and design constraints. It is characteristic of design tasks that designers often adjust their problem solving constraints in order to arrive at an agreeable solution (Lottaz and Smith 1997, Lyons 1995).

1.2.1 Task Description

The COCONUT task is related to those described in (Walker 1993, Whittaker et al. 1993) but differs in the emphasis and complexity of the task.[2] Each of the two participants in the task is given a separate budget and inventory of furniture that lists the quantities, colors, and prices for each item in that inventory.[3] Neither participant knows what is in the other's inventory or the money that the other has. The participants have the same types of knowledge but different instantiations of it. By sharing information, the participants can combine their budgets and can select furniture from each other's inventories. Purchasing decisions are joint; they must be mutually known and approved. The participants are equals in that there is no master-slave or expert-client relationship. Both participants have been briefed on the task goals, incentives and the tools and have had no prior contact.

[1] The remaining dialogues have not been fully annotated.

[2] Walker's similar task is performed by two artificial agents whereas our task and that in Whittaker et.al. is performed by two humans. Whittaker et.al.'s dialogues are spoken whereas ours are written.

[3] In Walker's task this information is committed to memory but in our task the participants have this information in written form.

The participants' main goal is to negotiate the purchases; the items of highest priority are a sofa for the living room and a table and four chairs for the dining room. The participants also have specific secondary goals which further complicate the problem solving task. Participants are instructed to try to meet as many of these goals as possible, and are motivated to do so by associating points with satisfied goals.[4] The secondary goals are: 1) Match colors within a room, 2) Buy as much furniture as you can, 3) Spend all your money.

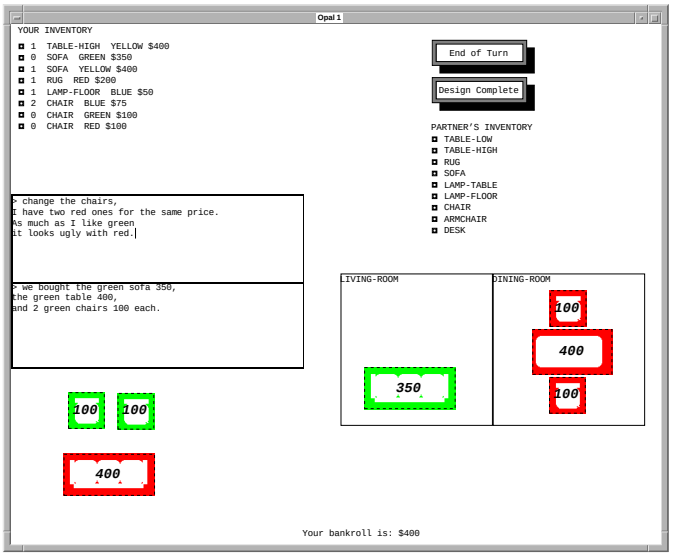

FIGURE 1 A View of the COCONUT Interface

1.2.2 Communications Setting

The participants are in separate rooms and can communicate via the computer interface only. They are asked to maintain private graphical representations of their discussions and incremental agreements. The participants share dialogue windows but the inventories, budgets and updated floor plans are private and appear only on the owner's color display. Figure 1 shows the interface as it looks in the middle of a design session.

The buttons in the upper right corner of Figure 1, "End of Turn" and "Design Complete", enforce turn-taking and initiate incremental

[4] In Whittaker et.al.'s task the incentives and goals are simpler.

recording of the conversation and the graphics updates. No interruption of the partner's turn is allowed. Also note that only the participants' current turns are available, i.e., the sender's current turn in the top dialogue box and the partner's previous turn in the bottom one.

During an incremental recording, the most recently transmitted message is recorded as well as the state of the sender's graphics display. The graphics display record is a description of the furniture icons in the two rooms as well as those that have been created but not assigned to any room. The participants incrementally update the floor plan by placing the furniture icons in meaningful locations. Whenever possible we have used this private information in our corpus analysis as partial evidence of what the speaker's utterance meant and what the hearer understood. However, the primary purpose of the graphics display is as a memory aid for the participants and is only intended secondarily to help clarify possible sources of misunderstanding during analysis.

Note that since a participant does not know what furniture his partner has available, there is a menu (see the mid-right section of the display in Figure 1) that allows a participant to define furniture icons that represent what he understands his partner to have as his partner shares this information with him. There is nothing to prevent the participant from creating an icon for a piece of furniture the partner does not actually have since the menu is general. An icon for a non-existent item could result from either a misunderstanding of his partner's item description or an error in selecting feature values for the item. At minimum the participant must know the type of the furniture item (e.g. chair, table). If the participant does not know or is uncertain about any of the other feature values of the furniture item, he can leave that feature unspecified (i.e. color and purchase price).

The participants first worked through a trial problem to familiarize themselves with the task and the communications setting. During this time they could ask for guidance on using the interface and clarification of the goals and incentives. The participants then solved 1-3 scenarios where the inventories and budgets vary. The problem scenarios ranged from ones where items are inexpensive and the budget is relatively large to ones where the items are expensive and the budget relatively small.

Nothing intrinsic to this task should result in unusual redescriptions. It is reasonable to assume that design tasks, and the COCONUT task in particular, should not affect the number of redescriptions. While we will see evidence that this specific task does lead to the inclusion of identificationally unnecessary attributes in redescriptions, we expect that this interaction should hold for a wide range of tasks where many object attributes are relevant to the problem solving task and where the definition

of success is also negotiable. However, because of the non-interruptible setting of the dialogues, attentional limits may also cause redescriptions to be longer than they would otherwise be (Garrod and Anderson 1987, Issacs and Clark 1987, Oviatt and Cohen 1991).

1.2.3 Estimating Overspecified Redescriptions

In our preliminary investigations of the COCONUT corpus, we noticed that there seemed to be a large number of overspecified redescriptions of furniture items. As we noted earlier, a redescription includes any explicit information in an utterance that could describe a discourse entity and that is mutually known. For example, if a shared discourse entity for a chair is mutually known to have the color *red*, then including *red* in the utterance, as with "My chair is red" makes it part of the redescription and if there is only one chair then it is also overspecified.[5] But if *red* is not a mutually known attribute then the redescription is defined as expressing only the type attribute *chair* and the owner attribute *self*.

We confirmed our initial impressions by first determining for each description what other furniture discourse entities might be salient for the dialogue participants. Following the terminology of (Dale 1992), we call these salient, mutually known entities, the distractors. Different definitions for a distractor set are suggested in the literature (Dale 1992, Passonneau 1996, Grosz and Sidner 1986, Levelt 1989, Krahmer and Theune 2001). Since it is not yet clear what definition of the distractor set is correct, we tried several plausible definitions that relate to current theories in computational linguistics and psycholinguistics. Using several definitions of the distractor set, we were able to see how many overspecified redescriptions resulted under each definition. Since there was a large proportion of overspecified redescriptions no matter what distractor set definition we used, we reasoned that the COCONUT corpus would be useful for studying our hypotheses. Below we give the details of how we identified overspecified redescriptions and of the distractor set definitions we tried. Finally we report the number of overspecified redescriptions we found.

To identify overspecified redescriptions we followed Passonneau's procedures for identifying overspecified noun phrases (Passonneau 1996). She used a distractor set that is the union of all the discourse entities (indicated by noun phrases only) in the current discourse segment (as indicated in (Grosz and Sidner 1986)) and all the entities in the last segment that most recently evoked the entity to be described. To be conservative, she assumes that if the most recent segment to evoke the

[5] Note that this example is also an informationally redundant utterance (Walker 1993).

target entity is not the same as the current segment then it is a resumption and the intervening focus spaces should not be included in the distractor set. The descriptive content that is needed to avoid ambiguity and the size of the distractor set are positively correlated. So Passonneau's model, which minimizes the distractors, will also provide a conservative measure of the number of overspecifications in a corpus. We will call this distractor set definition **SEG**.

Similarly to Passonneau, we first identified all the redescriptions that were potentially overspecified by selecting those that used more mutually known attributes than in their previous description. We then filtered this set of redescriptions using the **SEG** distractor set definition. With the first step, we found that 51% (84 of 166) of the redescriptions in the COCONUT corpus were potentially overspecified. And after filtering these with **SEG**, we found that 46% (76 of 166) of all the redescriptions were overspecified. This seems to confirm our initial impressions.

Because it is possible that **SEG** is not the best distractor set definition for all genres, we also tested for overspecification using some other cognitively motivated distractor set definitions. We followed the same methodology as above but we substituted the following distractor set definitions for **SEG**;

- **ALL**: all discourse entities previously mentioned in the discourse.
- **1UTT**: all discourse entities mentioned in the previous utterance.
- **SEG+**: all discourse entities in the current discourse segment and all the entities currently in the solution set
- **5UTT**: all discourse entities mentioned in the previous 5 utterances.

ALL and **1UTT** are two simple and rather implausible definitions for the distractor set and have been included to determine what happens at the extremes. Actually, there is some theoretical merit to **ALL**. (Poesio 1993) indicates that the distractor set should be a combination of the perceptual focus space and the discourse focus space. In the COCONUT setting, the designers often created graphics icons to help them remember the items their partner had described to them and which items they had presented to their partner. These graphical representations could behave as a perceptual focus space for the speaker. However, in view of the evidence discussed in (Clark and Marshall 1981), the participants would have to assume they are both keeping such a record (i.e. the representations would need to be part of their common ground) in order for the dialogue to proceed successfully.

1UTT, while extreme, also represents a focus space similar to that used in computational research on centering (Walker et al. 1997) to

	degree of redundancy	average distractor set size
SEG	46% (76)	5
ALL	39% (64)	19
1UTT	46% (76)	2
SEG+	46% (76)	4
5UTT	44% (73)	4

TABLE 1 Degree of redundancy under different distractor set definitions

determine acceptability conditions for anaphoric reference.

SEG+ assumes that items that have been selected already as part of the solution will remain in focus during the rest of problem solving. The reasoning task provides a rationale for this definition since these items serve to limit the money that is left to spend and may be applicable for determining color match constraints. Finally, **5UTT** is a simple approximation of recency in the discourse.

Table 1 shows the percentage of overspecified redescriptions under each distractor set definition. The degree of overspecification runs as low as 39% with the **ALL** distractor set definition and as high as 46% with the **SEG**, **1UTT** and **SEG+** definitions. No matter which of these distractor set definitions we use, there is still a high degree of overspecification in the COCONUT corpus.

1.3 Potential influences on redescriptions

Now we turn to the question of what else in addition to the identification goal could intentionally influence attribute selection for redescriptions. Our hypotheses reflect non-identification goals that are motivated by the functions of repetition at the utterance or propositional level (Walker 1993, Johnstone 1994) and the inferences that relate utterances to one another and make a discourse coherent (e.g. changes to the color match constraint that are not directly communicated by the dialogue participants).

The first type of task-related inference we considered is motivated by the observation that participants in task-oriented dialogues appear to be able to coordinate on the relaxation of particular task constraints without needing to discuss it. For example, the participants may decide it is impossible to achieve the optional task goal of matching furniture colors within a room. In the COCONUT dialogues, in 38% of the cases where optional goals were abandoned, the participants appeared to agree

to abandon the goal without explicit discussion.[6] Our hypothesis is that this inference can also be cued by the content of a redescription when it realizes attributes of a domain object that are not needed to identify which object is under discussion. For example, in (1)[7] A specifies both the color and price for both the sofa and the lamp even though the price attributes alone would adequately identify each item. By specifying the color, one can easily infer that the color match constraint has been dropped in the proposal. A has eliminated having to explicitly communicate this information (Walker 1993) but has reduced the risk of the hearer missing the inference (Carletta 1992).

(1) S: <...> if we do that i have 400 blue sofa and a 350 yellow sofa, and i have a 250 blue floor lamp or a 150 yellow rug. <...>

A: <...> so now we have 600 left for the living room. if we get *your 350 yellow sofa* and *your 250 blue floor lamp*, that sounds good to me because I don't have anything better in my inventory.

DOMAIN CONSTRAINT CHANGES HYPOTHESIS: Attributes related to constraint changes are expressed in a context where the change is to be inferred by the hearer.

The second task related inference is based on previous research that suggests that discourse relations between utterances, such as *motivation*, can influence the content and form of utterances (Mann and Thompson 1987, McKeown 1985, Moser and Moore 1995). It seems plausible that the speaker can cue these same inferences via redescriptions. For example, in (2) one can infer from O's last utterance and the redescription *mine for 150* that his motivation for proposing his rug is its better price.

(2) U: i have a blue rug for 250. that would leave us with 50 or any other options you may have for us.

O: ok lets take the blue rug for 250, my rug would not match which is yellow for 150.

U: we don't have to match...

O: well then lets use *mine for 150*.

PERSUASION HYPOTHESIS: Attributes that are relevant to getting the hearer to agree with the speaker's proposed action may be expressed in the context of a goal to propose that action.

The next two types of inference are based on the idea that if a speaker repeats an utterance and provides no new information, this can show

[6] In (2) there is some explicit discussion about the color match goal.

[7] All of the COCONUT excerpts appear verbatim except that we italicize redescriptions and omit parts of turns when they are unrelated to the point of the example. We indicate omissions with <...>.

that a stage of the interaction is complete (Whittaker and Stenton 1988, Jordan and Di Eugenio 1997). Repeating attributes for a recently evoked item could show that the current stage has just been completed while doing so for an older item could indicate that a higher level subproblem has been completed. In (3), S's second utterance appears to end a stage in the interaction, in this case the end of the agreement process for a *select sofa* action (Di Eugenio et al. 2000).

(3) S: <...> I have a $300 yellow sofa <...>
G: My sofa's are more expensive so buy *your $300 yellow sofa.* Also <...>
S: <...> *I* will go ahead and buy *the $300 yellow sofa.*

COMMITMENT HYPOTHESIS: In the context of a joint commitment to a proposal, all the attributes expressed in the proposal will be repeated.

The second case in which a higher level subproblem was completed is illustrated by the summary in (4). Note that D summarizes both living room (as requested) and dining room items. Summaries differ from commitments in that they are delayed redescriptions. The action associated with the object was completed and the participants had moved on to a new part of the task.

(4) G: I got the rug. What do you have in the living room and what are the prices of the items
D: the green sofa in the living room 350. dining room—> *3 yellow chairs 75 each, 1 high-table yellow, 1 yellow rug*

SUMMARIZATION HYPOTHESIS: In the context of a previously completed problem or subproblem, all the mutually known attributes for an item will be repeated.

The final type of inference we considered is when a speaker repeats an utterance to show that it was understood (Clark and Schaefer 1989, Brennan 1990, Walker 1992,1993). In the COCONUT corpus, the hearer sometimes repeats the description in the turn immediately following. For example, in (3) G repeats S's description of the sofa, although the sofa was introduced by S. We claim that this type of redescription could help verify that the attribute information was correctly understood.

VERIFICATION HYPOTHESIS: In the context of a newly introduced entity, all the attributes expressed will be repeated by the hearer in his/her next turn.

1.4 Analyzing the Corpus

To verify our hypotheses, we undertook a two part corpus investigation. First, we did correlational studies of the corpus using factors derived from annotation features related to the agreement process as described in (Di Eugenio et al. 2000), other discourse features, discourse entities and the problem solving state. However, since one cannot infer causality merely from correlational studies, in the second part of our investigation we analyzed the performance of parameterized attribute selection strategies which used the annotated corpus as input and output test data.

For both parts of the investigation, we needed to define the inference contexts we described in the hypotheses in terms of the annotated features and other easily extractable features of the corpus (e.g. the utterance speaker and the proximity of a redescription to its last mention). As part of recognizing the contexts, we used the following agreement process definitions[8] which were presented and supported in (Di Eugenio et al. 2000):

- propose: The speaker offers the item and unconditionally commits to using it and the offer makes the mutual solution state determinate.
- partner decidable option: The speaker offers an item and conditionally commits to using it but the offer leaves the mutual solution state indeterminate.
- unconditional commit: The speaker indicates his unconditional commitment to using the item
- unendorsed option: The speaker offers an item but does not show any commitment to using it when the mutual solution state is already determinate.

The annotation features that comprise these definitions were all found to have good intercoder reliability (Di Eugenio et al. 1998).

1.4.1 Annotation Scheme

We developed two additional types of corpus annotation features to support our study: (1) discourse entity level annotations that capture (a) the definitions and updates for discourse entities as a dialogue progresses and (b) the attributes selected to describe discourse entities, and (2) utterance level annotations that capture problem solving and discourse features.

[8]The definitions presented here are abbreviated. There is also an aspect that relates to the problem solving architecture which distinguishes them from speech acts.

Features of type (1) are needed to supply information about the furniture entities evoked in the dialogue. The main objective was to identify the entity being communicated and how the information about that entity was communicated. Both initial and subsequent references were annotated so that we could capture how the description of a single discourse entity developed during the course of the dialogue. By tracking the discourse entities in the dialogue we could tell when a subsequent reference to an entity might also add new information about the entity or correct erroneous information. For example in, "I have a $200 table. It is green.", entity_1 from the first utterance is ((type table)(owner A)(price 200)). The pronoun "it" in the next utterance corefers to entity_1 but the utterance also adds to it new information about the color of the object. The entity description then gets updated to ((type table)(color green)(owner A)(price 200)). These entity descriptions serve as input to the attribute selection strategies. However, the strategies cannot choose to use attributes that are new to an entity (i.e. not mutually known) to corefer. In comparing a strategy's selections with those made by the human, choices about whether to describe a new attribute are not considered.

For the furniture entities, we asked annotators to indicate the attribute-value pair information for each discourse entity in an utterance, and the sources for this information (e.g. from the utterance, the NP or locally inferred). Annotators were also asked to indicate whether the discourse entity was new or a coreference to a previous discourse entity and to what other discourse entities the current entity might be related. Here, some of the relevant relations include set, part-of, and class relations. Finally, we also asked the annotators to indicate the action for which the discourse entity was an argument.

In addition to the agreement process components we described earlier, we also needed other problem solving and discourse features at the utterance level to test our hypotheses. First, we needed to know what constraint changes were communicated, and whether these changes were communicated explicitly or implicitly. We assumed a set of initial constraint settings that would maximize the number of points earned. In general, these initial settings held true for all of our participants since the task instructions that explained the scoring for solutions was the only common ground that the participants had at the start of the problem solving trials. Annotators were instructed to pick an appropriate constraint description from a given list whenever there was a change to that constraint from its previous setting.

Finally, we also needed to identify the task structure and the discourse segments. We used a change to a different domain action as a

cue for the non-linguistic task structure (Terken 1985). Each domain action provides a discourse segment purpose so that each utterance that relates to a different domain action or set of domain actions defines a new segment.

We assumed that there were at least three and at most five component actions to be discussed (distinguished by furniture type and room): selecting four chairs for the dining room, selecting a table for the dining room, selecting a sofa for the living room, and selecting a set of optional items for the living room and dining room. We instructed the annotators to determine the actions addressed in each utterance by considering whether any furniture items or furniture templates (e.g. "do you have a red sofa?") being discussed in the utterance could unambiguously be related to one of these actions. Annotators were also asked to distinguish between when an action was first addressed and when the utterance continued the discussion. If the relation of the furniture item or template to actions was ambiguous, the annotators were instructed to indicate the highest level action that was unambiguous (e.g. select items for the dining room). Contiguous utterances that discussed a particular action were taken to define a discourse segment. Utterances that introduced or restarted action discussions while also continuing active discussions of other actions, were interpreted as starts of embedded discourse segments.

To develop and validate the annotation scheme, we conducted intercoder reliability studies using a balanced subset of the corpus. 30% of the corpus was annotated by two annotators for the purpose of determining intercoder reliability.[9] We use the Kappa coefficient of agreement (Krippendorff 1980, Carletta 1996) to assess intercoder reliability; this measure factors out chance agreement between coders. The discourse processing community uses Krippendorff's scale (Krippendorff 1980) to interpret and apply the Kappa coefficient, which varies between 0 and 1. Krippendorff's scale discounts any variable with $K < .67$, allows tentative conclusions when $.67 < K < .8$, and definite conclusions when $K \geq .8$. Table 2, which shows the intercoder reliability results after two development iterations and one reconciliation meeting that identified omissions, suggests that all of the features are defined clearly enough so that they can be reliably annotated and used in studies. After establishing the intercoder reliability, additional dialogues were annotated by one annotator. The overlap between the dialogues that were annotated for this study and those annotated by the COCONUT project for the agreement structure resulted in 13 fully annotated dialogues which we used

[9]One annotator's area of expertise is linguistics; the other is the author of this paper.

for testing our hypotheses about attribute selection in redescriptions.

Actions & Constraints	Introduce Actions	Continue Actions	Change Constraints	
	.897	.857	.881	
Discourse Entities	Reference Coreference	Discourse Relations	Attributes	Entities to Actions
	.863	.819	.861	.857

TABLE 2 Kappa values for the Annotation Scheme

1.5 Results of the Correlational Analysis

We used chi-square and the Fisher exact test[10] to check for correlations between factors in the corpus. The factors are all derived from the annotated features we described earlier. Although these tests assume independence, we feel we can violate this assumption given that the dependencies in a discourse aren't direct and obvious. In all of the contingency tables we will present, the counts are restricted to utterances that contain redescriptions. Finally the counts were all done automatically using software that interpreted the annotation features since the contextual factors generally involved multiple annotation features.

Domain Constraint Changes Hypothesis

For this hypothesis we test whether there is a difference in attribute usage when a constraint change is communicated implicitly or explicitly. Recall that COCONUT is directly annotated with features indicating (1) whether a constraint change was communicated and whether this was accomplished implicitly or explicitly (2) which attributes were included in the redescriptions.

Changes	Related Attributes
Room Color Limit	color
Price Upper Limit	price
Price Evaluator	price
Attribute Limit	color, price

TABLE 3 Associated Attributes and Changes

We only counted attributes that relate to constraints. For example, we only looked at the usage of the color attribute for the color match

[10]We use the Fisher exact test when $N < 20$ or an expected cell frequency is ≤ 5.

constraint or price for placing price limits. In Table 3, we list each of the constraint types that we examined and the attribute that we expected would be useful for inferring that change.[11] Our expectations derive from the instructions given to the COCONUT dialogue participants.

	Attribute Used	Attribute not Used
Implicit change	9	0
Explicit change	2	11

TABLE 4 Contingencies for Domain Constraint Changes Hypothesis

Table 4 shows that in the context of an implicit constraint change, attributes related to the change are more likely to be used in the description than when the change is explicit (Fisher Exact Test, $p < 0.0002$).

Persuasion Hypothesis

For the Persuasion hypothesis, we wish to test whether expressing an attribute in a redescription is related to whether the expressed attribute makes the redescribed item more desirable as a solution for an action than the alternatives. For example, the cost of the item being redescribed might be lower than any of the alternatives that have been discussed so far.

A persuasion context exists when a proposal is to be made and alternate solutions exist and there is a contrast between the colors or prices that make the proposed item clearly a better choice. Given the analysis of the agreement process in (Di Eugenio 2000), we first look for either a propose utterance, or an unconditional commitment utterance where the previous state is a partner decidable option, an unendorsed option or a list of options in which the speaker intentions are unclear.

For each of the unconditional commitment cases, we present examples. First, in (5), A's partner decidable option is followed by B's unconditional commitment.

(5) A: I have a blue sofa for $200.
B: I have a yellow sofa for $250. Let's go with your $200 sofa.

In (6), B does not endorse the option he presents but A overrides his objection with an unconditional commitment to it.

(6) A: We have $100 left. I still have that $50 blue chair.
B: I have a rug for $100, but it is yellow.

[11] The relevant attribute for the attribute limit constraint is indicated in the annotation for the constraint change.

A: We don't need to match. Let's get your $100 rug.

Finally, in (7), A lists all of the items he has available. From the perspective of the agreement structure, lists such as this have no action intention associated with them. However, the items do become part of the dialogue participants shared knowledge allowing all the items to be considered during problem solving so that they can become alternative options for the actions they are implicitly associated with. Because of this, B's second utterance is annotated as an unconditional commitment because he was in a position to deliberate.[12] In this case there are two possibilities for what sofa to select and so a persuasion context arises.

(7) A: I only have 2 red tables for $200, 1 green table for $350 and 4 $50 blue chairs. I don't have any rugs or lamps but I have 1 yellow sofa for $200.

B: I have yellow rug for $75 and a blue sofa for $200. Let's buy your yellow sofa and my rug.

Once we have identified possible proposals, we need to check for contrasts with alternatives. The alternatives are approximated by accumulating a list of the items evoked for each action. After a propose or unconditional commitment, all the items in the list for an action get flushed before starting over with the proposed item. The list must still be maintained after a proposal in case a counterproposal arises.

Contrast	Related Attribute
Matches room but not alternatives	color
Cheaper than alternatives	price
More expensive than alternatives (near end of problem)	price

TABLE 5 Associated Attributes and Contrasts

Next we check for contrasts between the item being proposed and the alternatives.[13] The contrast possibilities are shown in table 5 and arise from the COCONUT problem description. We were unable to accurately model the goal of buying as many items as possible with the annotations available. For color we compare the color of the proposed item to those items already selected for the room and the alternative items. If the

[12] This requirement for unconditional commitment is related to the problem solving architecture and is justified in (Di Eugenio 2000).

[13] (Krahmer and Theune 2001) also check for contrastive attributes in spoken dialogue applications.

proposed item matches items already selected for the room while none of the alternatives do, then a persuasion context exists. For prices there are two possibilities that depend on whether or not the end of the problem solving effort is nearing. An item may be a better choice 1) when the price of the proposed item is greater than that of each alternative (i.e. it may be helping to spend out the budget) or 2) when the price of the proposed item is less than that of each alternative (i.e. the cheaper item may be preferred since it leaves some money for other purchases).

	Attribute Not Used	Attribute Used
no contrast	18	9
contrast	13	24

TABLE 6 Contingencies for Persuasion Hypothesis

Table 6 indicates that in contexts where a contrast is predicted, the contrastive attribute is more likely to be included in a redescription ($\chi^2 = 5, p < .05, df = 1$).

Commitment Hypothesis

Here we test whether in the context of a joint commitment to a proposed action all the attributes expressed in the proposal are more likely to be repeated. A joint commitment context exists when either 1) there is a previous proposal or unconditional commitment for the action involving the entity in the immediately previous turn and no other items have been discussed for the action in the interim or 2) a speaker unconditionally commits again after doing so in his previous turn.

When determining repeated attributes, we discount the type and owner attributes. The type attribute is excluded because it involves pronominalization and zero anaphora; issues we are not addressing in this research. We exclude the owner attribute because its only function is identification in this domain.

	Not Repeat Attributes	Repeat Attributes
No Commitment	7	8
Commitment	2	20

TABLE 7 Contingencies for Commitment Hypothesis

Table 7 indicates that in contexts where a joint commitment is predicted, all mutually known attributes are more likely to be included in

redescriptions (Fisher Exact Test, $p < .0171$).

Summarization Hypothesis

Here we test if the previous completion of a problem or subproblem is related to the expression of all the mutually known attributes in a redescription. First, we must isolate redescriptions that occur after an agreement has been reached for the action.

A summarization context exists when an agreement has been reached for the action without the action being readdressed between the agreement and the current turn. The achievement of an agreement state is approximated when either 1) a propose or partner decidable option was the last state for the action and it happened more than two turns ago or 2) an unconditional commit was the last state and it happened two or more turns ago. In the first case, the agreement must be inferred and in the other the agreement is more explicit.

For the agreement state under condition 1), we require more than two turns to intervene because we want to allow for the cases where the partner left the decision pending by moving on to a dependent action (e.g. a final table decision may be left pending until the chair options are explored). We are estimating that if the action is not revisited after three turns, then it was not put on hold pending work on another action and that the partner agreed by moving on to another independent action.[14] This test for agreement takes into consideration that the initiation of the relevant next contribution shows evidence of understanding (Clark and Schaefer 1987) and possibly joint commitment. For condition 2), we require that there be an intervening turn so that the partner is able to show that he has moved on to some other problem.

	All Mutual Attributes Used	Not All Mutual Attributes Used
Not End of Agreement Process	54	117
End of Agreement Process	8	8

TABLE 8 Contingencies for Summarization Hypothesis

As with the commitment hypothesis, the type and owner attributes are excluded when determining whether mutually known attributes are

[14] In the initial version of the annotation scheme, there was a feature for indicating dependent actions but it was dropped because of poor intercoder reliability.

repeated.

Table 8 indicates there is no correlation between a summarization context as we have characterized it and whether all the mutually known attributes that relate to decisions get repeated ($\chi^2 = 1.49, df = 1$, NS).

Verification Hypothesis

	Attributes Not All Repeated	Attributes All Repeated
initial not in previous turn	1	0
initial in previous turn	44	2

TABLE 9 Contingencies for Verification Hypothesis

With this hypothesis we test whether the repetition of all the attributes presented in a previous description correlate with a context in which the entity was just introduced. In this case we collect all the attributes that were presented in the turn where the item was first described and check whether this mention of the item was in the immediately previous turn or further back in the dialogue. As with the commitment and summarization hypotheses, the type and owner attributes are excluded when determining whether attributes are repeated. Table 9 shows no correlation between the verification context and the choice of attributes ($\chi^2 = .06, df = 1$, NS).

1.6 Comparing Redescription Strategies

The first part of our study shows which of the contexts that predict communicative goals are more likely to influence redescriptions—the inference contexts and attribute choices indicated in our hypotheses positively correlated for all but Verification and Summarization. In what follows, we will describe our experimental comparisons of selection strategies for redescribing objects—two identification-only strategies and a strategy that incorporates identification and responds to the contexts indicated in our hypotheses.

We analyzed how well computer simulated selections for the COCONUT corpus matched human selections. We reasoned that if our hypotheses were valid then a selection strategy that incorporates them should match the selections made by humans at least as well as an identification-only selection strategy. We anticipated that the degree of match to humans could be similar between the selection strategies since

there may be many allowable ways to express a description for identification purposes and the selections intended to cue the inferences could intersect some of these allowable ways.

We used the human generated descriptions in the COCONUT corpus to evaluate the descriptions created by the selection strategies we wished to test. We simulated selections for the COCONUT dialogues by using annotations about the discourse entities to be evoked and the contexts in which they appeared as input to the selection strategies. To compare the performance of a selection strategy to that of humans, we used a measure of the degree of match between the human's and the parameterized strategy's selection of attributes for the same discourse entity in the same dialogue context. Inclusion and exclusion of an attribute both count in the degree of match. A perfect match means that the strategy included or excluded the same attributes as the human did for a particular entity. The measure, X/N, ranges between 0 and 1 inclusive, where X is the number of attribute inclusions and exclusions that agree with the human data and N is the number of attributes that can be expressed for an entity. This response variable is called *match* in the experiments that follow.

To determine the best internal parameter settings for each strategy and to compare strategies we first did an analysis of variance (MathSoft Inc. 1998) on the results of the experiments. The analysis of variance indicates whether there were any significant differences in the performance as we varied the parameter settings or redescription strategy. To determine where and how large any performance differences are, we then did either multiple pairwise comparisons (MCA) (Hsu 1996) or multiple comparisons with a control (MCC) (Dunnett 1964).[15] We display the results of the multiple comparisons as 95% confidence intervals, (e.g. as in Figure 2), and they are always of the form:

$$(\text{estimate}) \pm (\text{critical point}) \times (\text{standard error of estimate})$$

The critical point in the above calculation depends on the multiple comparison method used (e.g. Tukey, Dunnett, LSD). We chose the method that created the smallest critical point and the selected method is indicated in each graph.[16]

Intervals in the graphs that exclude zero indicate statistically significant performance differences. The labels on the y axis indicate the two levels or experimental factors that were compared and represent the

[15] We used S-plus' multicomp function to perform the multiple comparisons (MathSoft Inc. 1998).

[16] S-plus' multicomp function can optionally consider all the the valid methods to find the smallest critical point.

differences in performance. If the interval is to the right of zero then the first member of the label pair performed better and if the interval is to the left then the second member of the pair performed better.

It doesn't follow automatically that performance is identical when there are no significant differences. We will discuss non-significant differences in terms of performance trends in making judgement calls about equality of performance. If the center point of an interval is to the right of zero then we will say that the first member of the label pair has a trend towards performing better, and vice versa if the center point is to the left of zero.

1.6.1 Defining the Attribute Selection Strategies

There are a variety of strategies suggested in the literature for satisfying the identification goal but many aspects of the strategies or the information they depend upon are vague. We will examine two different strategies for choosing attributes to satisfy the identification goal. The first strategy is the incremental algorithm (**INC**) described in (Dale and Reiter 1995). **INC** incrementally builds a description by checking an ordered list of attribute types and selects an attribute only when it rules out any remaining distractors. As distractors are ruled out, they no longer influence the selection process. The initial set of distractors are computed according to what is expected to be in focus for the speaker and the hearer based on the intentional structure of the dialogue.

The second strategy we examined for satisfying the identification goal is based on the gestalt search template (**gestalt**) described in (Levelt 1989). In this strategy, the template is overspecified in a way that makes the search for the referent easier. Following (Levelt 1989), we identified which static attribute template would maximize the number of redescriptions matched in the corpus and used it to create the base description for any entity that is to be re-evoked. We then supplemented the description using **INC** to rule out any remaining distractors.

These two selection strategies are parameterized for many lower level calculations which are not yet well specified by any theories but here we will only discuss determining the best distractor set definition to use within each selection strategy. The distractor set is used to assess whether a description will uniquely select the target object from its set of potential distractors.[17]

We will compare the performance of these two identification only selection strategies to each other and to a parameterized selection strategy called *intentional influences* (**IINF**). **IINF** tests for the inference con-

[17] (Krahmer and Theune 2001), in this volume, provide a more detailed specification of the distractor set. Our future experiments will incorporate their findings.

texts described in our hypotheses and selects appropriate attributes for each context that it finds in the corpus for the discourse entity that is to be redescribed. Afterwards, **IINF** then incrementally selects additional attributes as needed to rule out any remaining distractors. The identification strategy used when none of the inference contexts applies is parameterized so that we can incorporate the best identification strategy into **IINF**. **IINF** is also parameterized for which contexts are allowed to influence attribute selection so that we can determine which combinations of our hypotheses result in the best match to human descriptions.

1.6.2 Determining Internal Parameter Settings for the Selection Strategies

To find the best distractor set definition for both of the identification only strategies we varied only the discourse partitioning approach. There are many theories for partitioning the discourse but no empirical studies that conclusively support one over another. We will use the distractor set definitions we introduced earlier when estimating the degree of redundancy in the COCONUT redescriptions. Recall that we described two interpretations of partitioning that are based on the discourse segment purpose (Grosz and Sidner 1986), **SEG** and **SEG+**, and three extremes that assume no partitioning of the discourse other than recency, **5UTT**, **1UTT** and **ALL**.

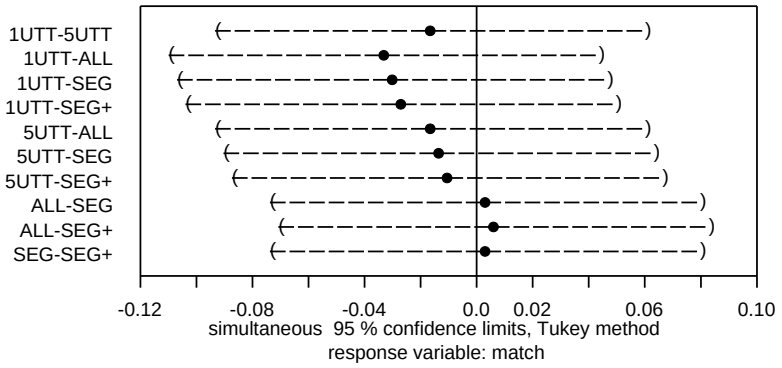

FIGURE 2 Comparing Distractor Set Definitions for INC

We found that while there were no significant differences between these definitions for **INC**, **ALL** was the better choice as shown in Figure 2. Likewise there were no significant differences between the definitions for **gestalt** but the more widely assumed and more restrictive

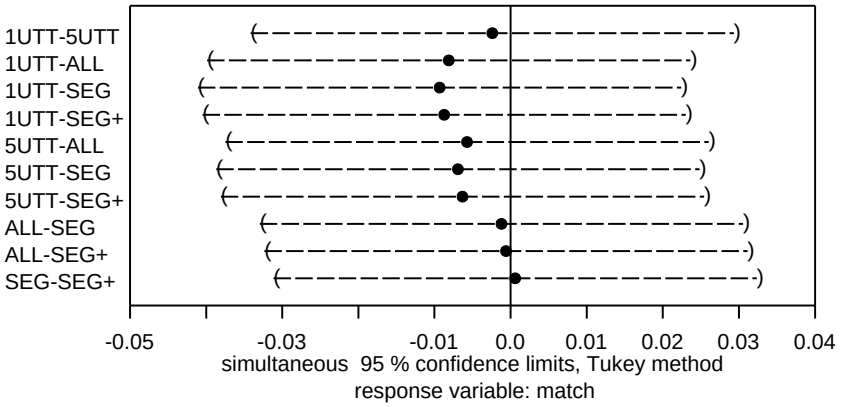

FIGURE 3 Comparing Distractor Set Definitions for Gestalt

definition, **SEG**, worked better as shown in Figure 3.

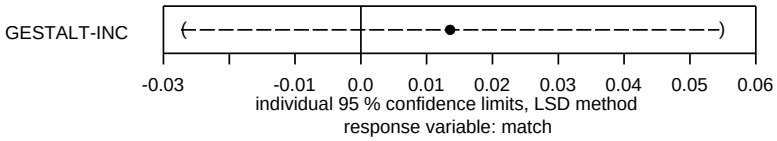

FIGURE 4 Comparing Identification Strategies

We found that when using the best distractor set definitions for each redescription strategy, the **gestalt** strategy, while not significantly better, had a trend towards better performance than the well known **INC** strategy as shown by the confidence interval in Figure 4. The main conceptual difference between the best versions of **gestalt** and **INC** was the gestalt search template. **INC** always includes the "type" attribute whereas the best **gestalt** setting was to always include both "type" and "color" attributes.[18] This means the main difference was that **gestalt** always included "color" in every redescription.

These results indicate that including "type" and "color" and not just "type" as with **INC** was a better strategy since it both performed slightly better and agreed with a standardly accepted way of partitioning a discourse (i.e. partitioning according to discourse segment purpose as

[18]Experiments to determine the best search template for **gestalt** are reported in (Jordan 2000).

with (Grosz and Sidner 1986)). Perhaps **INC** may have needed a reason to include "color" and a large distractor set would be more likely to justify the inclusion.

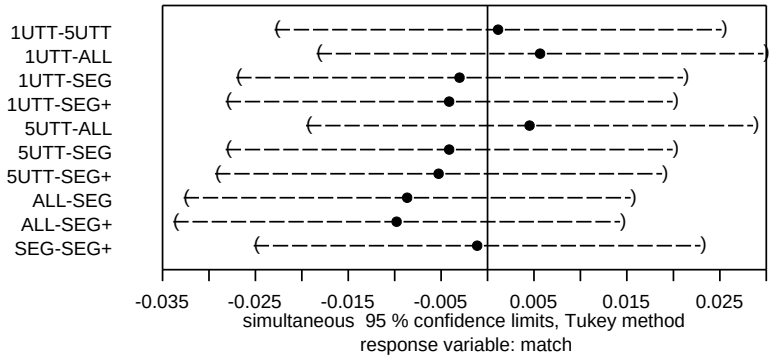

FIGURE 5 Comparing Distractor Sets for INC within IINF

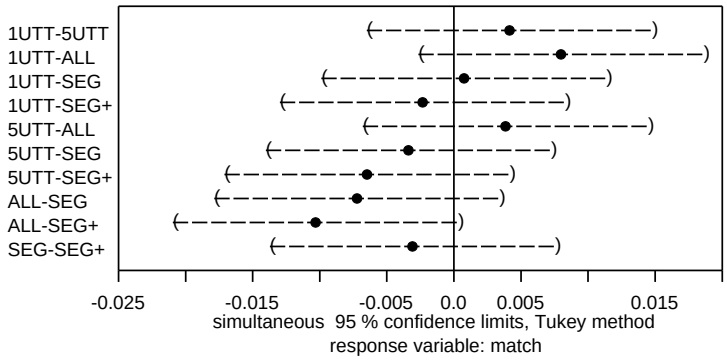

FIGURE 6 Comparing Distractor Sets for gestalt within IINF

After establishing the best internal parameter settings for the two identification strategies, we next did the same for **IINF**. It was parameterized for which contexts to check and for which identification strategy to incorporate. However, we did not want to assume that we should use the same settings when an identification strategy was embedded within **intentional influences** as when it was a stand-alone strategy that ignored other influences. Looking again at the distractor set definition

when identification was one of many possible goals that could influence attribute choices, we found that there were no significant performance differences for embedded **INC** or **gestalt**. But we can see from Figures 5 and 6 there is now a trend for the more restrictive **SEG+** to be the best setting for both.

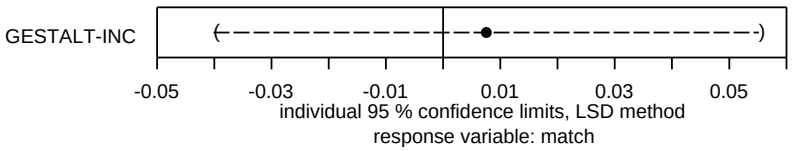

FIGURE 7 Comparing Identification Strategies within IINF

Finally, we compared versions of **intentional influences** using the newly determined settings for the incorporated identification strategy. There were no significant differences in performance, as shown in Figure 7, but there was still a trend towards better performance when **gestalt** was the embedded identification strategy. However, we can also see from this that when we are able to include other influences in addition to identification, the performance of the more widely used **INC** strategy tends to improve. This lends some additional credibility to our claim that there are multiple influences because, when we account for these influences, the standard theories and approaches tend to make choices more like that of humans.

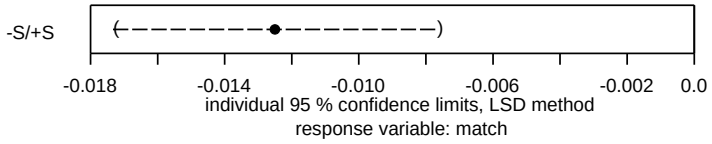

FIGURE 8 Comparing Summarization for IINF

Finally, by parameterizing the inference contexts that were considered by **intentional influences**, we were able to see which of our hypothesized contexts affected the performance of the **intentional influences** strategy. We accepted the positive correlational results from the first part of our study and only skeptically tested the negative ones. We found that Summarization had a clear positive influence while Verification had a clear negative one. For Summarization there is a signifi-

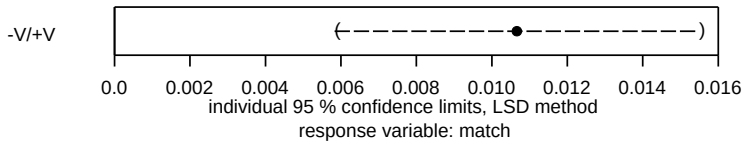

FIGURE 9 Comparing Verification for IINF

cant difference in performance ($F = 25.71, p < .0000004$) and the MCA comparison shown in Figure 8 indicates that it is better to include the summarization hypothesis. For Verification there is also a significant difference in performance ($F = 18.71, p < .00002$) but Figure 9 indicates it is better not to consider Verification.

In the case of the Summarization hypothesis, the results of our two part analysis may mean that our definition of the summarization context needs refinement. But with the Verification hypothesis, the results confirm it is not a valid influence for our corpus.

1.6.3 Comparing the Selection Strategies for Redescriptions

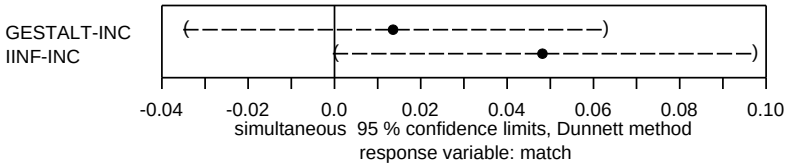

FIGURE 10 Comparing Redescription Strategies

Overall, we found that while the **intentional influences** strategy was not significantly better ($F = 2.58, p < .08$), it had a trend towards better performance compared to the stand-alone identification strategies, as shown by the MCC confidence intervals in Figure 10. The **intentional influences** strategy indirectly allows domain and discourse goals to override the pre-ordered list of attributes that represent attribute saliency in **INC**. For example, when a Persuasion context exists, it can allow attributes such as price to sometimes become more salient than color which is usually considered highly salient.

Although we do not have enough data to determine which, if any, of the contexts in our hypotheses is most influential for attribute selection, we can show in Table 10 the relative contributions of each of

Hypothesis	Percentage Contribution to Descriptions
Identification	29.33%
Commitment	26%
Summarization	22.67%
Persuasion	16.67%
Domain constraint changes	5.33%

TABLE 10 Contributions of Goal Contexts to Redescriptions

these contexts and the contribution of the identification goal within the **intentional influences** strategy. This gives us an informal view of the relative contributions without considering the frequency with which a particular context arises. The contribution made by the identification goal includes both the cases in which identification was the only predicted goal and the cases in which additional attributes had to be added to ensure unique identifiability after the initial selections made by the context checks. Although the contribution is smaller than one might expect, this does not mean that the identification goal was invalid for some redescriptions. Instead it indicates that the problem of identification was addressed already by the attributes that had been selected by the inference contexts. This reflects the economy that can be achieved with goal overloading.

1.7 Discussion

In comparing the performance of the **intentional influences** strategy to that of the two identification-only strategies, it is clear that there is a trend towards better performance especially in comparison to the incremental algorithm **INC**. We expected that the performance measures from the computer simulations could be statistically similar because there may be multiple allowable ways of redescribing some of the objects and the identification-only and **intentional influences** selection strategies could both be intersecting similar sized subsets of the actual expressions in the corpus.

An interesting question that arises in using human performance as the ideal measure is how well humans would agree with one another if asked to describe a particular entity in the context given in the corpus. (Yeh and Mellish 1997) found evidence that there are multiple possible solutions when deciding whether to use a zero anaphor, pronominal, full nominal, or nominal with just the head noun to redescribe entities. There was low agreement between human subjects about which form to use in

a set of test texts (Kappa of .41). Although we examined a subset of this issue, it is reasonable to expect that we would obtain similar results.

If we look at the actual mean matches of the strategies in Table 11 ($F = 26, p < 0$)[19], they fall far below the ideal. To give a bound on poor performance, we included the **RANDOM** strategy which loops over the number of attribute choices possible and randomly selects an attribute value to include.

Strategy	Mean Match
IINF	.6958
gestalt	.6611
INC	.6476
RANDOM	.4970

TABLE 11 Mean Strategy Performances

The argument for multiple solutions could mean that we have topped out on the performance measure. There could be speaker preferences with respect to the degree and type of overloading attempted. If this is true, we would also expect these preferences to change as the partners get acquainted and adapt to one another given the principle of LEAST COLLABORATIVE EFFORT (Clark and Wilkes-Gibbs 1986). Preferences and adaptation imply that we would always fall short of perfect agreement since we are looking at data from more than one speaker pair in our measures. In that case, we may be as close in agreement with the humans represented in the COCONUT corpus as other humans would be. If we are near the top line for performance, then we would not expect to see significant performance differences.

1.8 Conclusions

We have found evidence through corpus analysis and computer simulations that redescriptions do get overloaded with more than just the identification goal. We expect the types of inferences we have considered here to apply to other corpora but that the extent to which they apply depends on the underlying task and the communications setting. For example, there may be no problem solving constraints in the task that can be changed and so the selection strategy would never encounter a context in which this inference would be expected. Likewise, there may be inferences that are relevant for other corpora that did not appear in the one we analyzed. For instance, we would expect that something

[19]Recall that to agree with a human, $match = 1$.

like the verification inference might be more applicable in a face to face dialogue where information is more likely to be misheard or missed.

Clearly the extent to which we can study inferential influences on redescriptions will be limited by the complexity of the task that is the topic of the corpus. For example, it may be difficult to enumerate and annotate all the features that would signal that a particular inference is expected. However in a generation application in which we have access to the problem solving state there would not be the same problem and we could instead evaluate the comprehension effects of allowing different types of overloading in redescriptions. The results of our experiments should provide guidance as to what types of overloading would be most fruitful to try in applications that interact with users.

References

Appelt 1985. D. E. Appelt: Planning English referring expressions. *Artificial Intelligence* 26(1), pp.1-33.

Brennan 1990. S. E. Brennan: *Seeking and Providing Evidence for Mutual Understanding.* PhD thesis, Stanford University Psychology Dept., Unpublished Manuscript.

Carletta 1996. J. Carletta: Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2), pp.249-254.

Carletta 1992. J. C. Carletta: *Risk Taking and Recovery in Task-Oriented Dialogue.* PhD thesis, Edinburgh University.

Clark and Marshall 1981. H. H. Clark and C. R. Marshall: Definite reference and mutual knowledge. In A. Joshi, B. Webber, and I. Sag, editors, *Linguistics Structure and Discourse Setting.* Cambridge University Press, Cambridge, England, pp.10-63.

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

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

Clark and Wilkes-Gibbs 1986. H. H. Clark and D. Wilkes-Gibbs: Referring as a collaborative process. *Cognition* 22, pp.1-39.

Dale 1992. R. Dale: *Generating Referring Expressions.* ACL-MIT Series in Natural Language Processing. The MIT Press.

Dale and Reiter 1995. R. Dale and E. Reiter: Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science* 19(2), pp.233-263.

Di Eugenio et al. 1998. B. Di Eugenio, P. W. Jordan, J. D. Moore, and R. H. Thomason: An empirical investigation of collaborative dialogues. In *ACL-COLING98, Proceedings of the Thirty-sixth Conference of the Association for Computational Linguistics*, Montreal, Canada.

Di Eugenio et al. 2000. B. Di Eugenio, P. W. Jordan, R. H. Thomason, and J. D. Moore: The agreement process: An empirical investigation of human-human computer-mediated collaborative dialogues. *International Journal of Human-Computer Studies* 53(6), pp.1017-1076.

Dunnett 1964. C. Dunnett: New table for multiple comparisons with a control. *Biometrics* 20, pp.482-491.

Garrod and Anderson 1987. S. Garrod and A. Anderson: Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition* 27, pp.181-218.

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

Grosz and Sidner 1986. B. J. Grosz and C. L. Sidner: Attentions, intentions and the structure of discourse. *Computational Linguistics* 12, pp.175-204.

Heim 1983. I. Heim: File change semantics and the theory of definiteness. In R. Bauerle, C. Schwarze and A. von Stechow, editors, *Meaning, Use, and the Interpretation of Language.* Walter de Gruyter, Berlin.

Hsu 1996. J. C. Hsu.: *Multiple Comparisons: Theory and Methods.* Chapman and Hall, London.

Issacs and Clark 1987. E. A. Issacs and H. H. Clark: References in conversation between experts and novices. *Journal of Experimental Psychology: General* 116, pp.26-37.

Johnstone 1994. B. Johnstone: Repetition in discourse: A dialogue. In B. Johnstone, editor, *Repetition in Discourse: Interdisciplinary Perspectives, Volume 1*, volume XLVII of *Advances in Discourse Processes*, chapter 1. Ablex.

Jordan 2000. P. W. Jordan: *Intentional Influences on Object Redescriptions in Dialogue: Evidence from an Empirical Study.* PhD thesis, Intelligent Systems Program, University of Pittsburgh.

Jordan and Di Eugenio 1997. P. W. Jordan and B. Di Eugenio: Control and initiative in collaborative problem solving dialogues. In *Computational Models for Mixed Initiative Interaction. Papers from the 1997 AAAI Spring Symposium. Technical Report SS-97-04*, The AAAI Press, pp.81-84.

Kamp 1993. H. Kamp: *From Discourse to Logic; Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory.* Kluwer Academic Publishers, Dordrecht Holland.

Karttunen 1976. L. Karttunen: Discourse referents. In McCawley, J., editor, *Syntax and Semantics*, volume 7. Academic Press.

Krahmer and Theune 2001. E. Krahmer and M. Theune: Efficient context-sensitive generation of referring expressions. In K. van Deemter and R. Kibble, editors, *Information Sharing* (eds). CSLI Publications.

Krippendorff 1980. K. Krippendorff: *Content Analysis: an Introduction to its Methodology.* Beverly Hills: Sage Publications.

Levelt 1989. W. J. M. Levelt: *Speaking: From Intention to Articulation.* MIT Press.

Lottaz and Smith 1997. C. Lottaz and I. Smith: Collaborative design using constraint solving. From Swiss Workshop on Collaborative and Distributed Systems, Lausanne Switzerland.

Lyons 1995. K. W. Lyons: Collaborative design for assembly of complex electro-mechanical products. Presentation abstract for NCMS Manufacturing Technical Conference.

Mann and Thompson 1987. W. Mann and S. Thompson: Rhetorical Structure Theory: A Framework for the Analysis of Texts. Technical Report RS-87-190, USC/Information Sciences Institute.

MathSoft Inc 1998. MathSoft Inc.: *S-Plus 5 for Unix Guide to Statistics*. MathSoft, Inc., Seattle, Washington.

McKeown 1985. K. R. McKeown: *Text Generation. Using Discourse Strategies and Focus Constraints to Generate Natural Language Text*. Cambridge University Press.

Moser and Moore 1995. M. Moser and J. D. Moore: Investigating cue placement and selection in tutorial discourse. In *Proceedings of 33rd Annual Meeting of the Association for Computational Linguistics*, pp.130-135.

O'Donnell et al. 1998. M. O'Donnell, H. Cheng and J. Hitzeman: Integrating referring and informing in NP planning. In *COLING-ACL'98, Workshop on Computational Treatment of Nominals*.

Oviatt and Cohen 1991. S. L. Oviatt and P. R. Cohen: Discourse structure and performance: Efficiency in interactive and non-interactive spoken modalities. *Computer Speech and Language*, pp.297-326.

Passonneau 1996. R. J. Passonneau: Using centering to relax Gricean informational constraints on discourse anaphoric noun phrases. *Language and Speech* 39(2-3), pp.229-264.

Poesio 1993. M. Poesio: A situation-theoretic formalization of definite description interpretation in plan elaboration dialogues. In P. Aczel, D. Israel, Y. Katgiri, and S. Peters, editors, *Situation Theory and its Applications*, volume 3, pp.339-374. CSLI Publications.

Pollack 1991. M. E. Pollack: Overloading intentions for efficient practical reasoning. *Noûs* 25, pp.513-536.

Stone and Webber 1998. M. Stone and B. Webber: Textual economy through close coupling of syntax and semantics. In *Proceedings of 1998 International Workshop on Natural Language Generation*, Niagra-on-the-Lake, Canada.

Terken 1985. J. M. B. Terken: *Use and Function of Accentuation: Some Experiments*. PhD thesis, Institute for Perception Research, Eindhoven, The Netherlands.

Vonk et al. 1992. W. Vonk, L. G. M. M. Hustinx and W. H. G. Simons: The use of referential expressions in structuring discourse. *Language and Cognitive Processes* 7(3-4), pp.301-333.

Walker et al. 1997. M. Walker, A. Joshi, and E. Prince: Centering in naturally occurring discourse: An overview. In M. Walker, A. Joshi, and E. Prince, editors, *Centering Theory in Discourse*, Oxford University Press, Oxford. pp.1-28.

Walker 1992. M. A. Walker: Redundancy in collaborative dialogue. In *Fourteenth International Conference on Computational Linguistics*, pp.345-351.

Walker 1993. M. A. Walker: *Informational Redundancy and Resource Bounds in Dialogue*. PhD thesis, University of Pennsylvania.

Webber 1978. B. L. Webber: *A Formal Approach to Discourse Anaphora*. PhD thesis, Harvard University. Garland Press.

Whittaker et al. 1993. S. Whittaker, E. Geelhoed and E. Robinson: Shared workspaces: How do they work and when are they useful? *International Journal of Man-Machine Studies* 39, pp.813-842.

Whittaker and Stenton 1988. S. Whittaker and P. Stenton: Cues and control in expert client dialogues. In *Proc. 26th Annual Meeting of the ACL, Association of Computational Linguistics*, pp.123-130.

Yeh and Mellish 1997. C. L. Yeh and C. Mellish: An empirical study on the generation of anaphora in Chinese. *Computational Linguistics* 23(1), pp.169-190.