Can Nominal Expressions Achieve Multiple Goals?: An Empirical Study

Pamela W. Jordan
Intelligent Systems Program
University of Pittsburgh
Pittsburgh PA 15260 *
jordan@isp.pitt.edu

Abstract

While previous work suggests that multiple goals can be addressed by a nominal expression, there is no systematic work describing what goals in addition to identification might be relevant and how speakers can use nominal expressions to achieve them. In this paper, we first hypothesize a number of communicative goals that could be addressed by nominal expressions in task-oriented dialogues. We then describe the *intentional influences* model for nominal expression generation that attempts to simultaneously address the *identification* goal and these additional goals with a single nominal expression. Our evaluation results show that the *intentional influences* model fits the nominal expressions in the COCONUT corpus as well as previous accounts that focus solely on the *identification* goal.

1 Introduction

Previous work on nominal expression generation has mainly focused on the use of nominal expressions to achieve a speaker's goal to *identify* an object in the discourse context (Dale and Reiter, 1995; Passonneau, 1995). While other work suggests that it should be possible for a nominal expression to contribute to the satisfaction of additional goals (Appelt, 1985; Pollack, 1991; Stone and Webber, 1998), there is no systematic work describing what these goals might be and how speakers can use nominal expressions to achieve them. For example, consider the dialogue contribution in (1) in a context in which the color of the table is not necessary to IDENTIFY the discourse entity under discussion, but where the color of the item could be inferred to be MOTIVATION for the proposal.[1]

(1) Let's use my table. It is red.

A plausible hypothesis is that the alternative utterance in (2) could also support the MOTIVATION inference, and that the nominal expression *my red table* could thus simultaneously achieve the two goals of *identifying* the object under discussion and supporting the MOTIVATION inference.

(2) Let's use my red table.

We hypothesized that in addition to the conversational inference of MOTIVATION that a speaker might also attempt to achieve other task-relevant inferences via the generation of nominal expressions.

In order to test this hypothesis further, we first specified a number of specific communicative goals that we believe can be addressed by nominal expressions in task-oriented dialogues. We then implemented two models of nominal expression generation. The first model was the INCREMENTAL MODEL of Dale and Reiter (1995), which implements a strategy for satisfying the *identification* goal.

*This work was partially funded by NSF grant IRI-9314961 at the University of Pittsburgh, and a Mellon Predoctoral Fellowship

[1] For example, a possible context is one in which the speaker only has one table, and the speaker and hearer are trying for one color in a room, and have already agreed upon other red furniture.

The second model we call the *intentional influences* model; this model of nominal expression generation attempts to simultaneously achieve the *identification* goal and to cue other task-related inferences with a single nominal expression. We then evaluated these two models using 13 of the dialogues from the COCONUT corpus of task-oriented dialogues (Di Eugenio et al., 1998). This subset of dialogues contains 166 non-pronominal discourse anaphoric expressions which we call REDESCRIPTIONS (e.g. *my table* and *my red table* in (1) and (2)). Our results show that the *intentional influences* model fits the redescriptions in the corpus as well as previous accounts which focus solely on the *identification* goal. To test the validity of the set of inferences we considered in the *intentional influences* model, we also compared it to a model that addresses the identification goal but also includes additional mutually known attributes at random. We found that the inference set we considered was significantly better than random.

Section 2 describes the COCONUT corpus, the task that the conversants were attempting to achieve, the conversational inferences relevant to the task, and our hypotheses about the way these conversational inferences could provide additional influences on the speaker's generation of nominal expressions. Section 3 explains how we tested our hypotheses and section 4 presents our results.

2 Potential Influences on Redescriptions

The COCONUT corpus consists of 24 computer-mediated dialogues in which two people collaborate on a simple design task, buying furniture for two rooms of a house. 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 constrain 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. The secondary goals are: 1) use one color attribute value for all items within a room, 2) buy as much furniture as you can, 3) spend all your money. The participants are equals and must agree on the final plan for furnishing the house.

We hypothesized that many of the task-related inferences that the participants must make in this domain to (1) efficiently come to an agreement, and (2) do well on the task, could potentially be cued by the nominal expressions describing the items of furniture used to solve the task.[2]

Persuasion Hypothesis: The first task-related inference that we consider is the MOTIVATION inference exemplified in examples (1) and (2). Previous research suggests that discourse relations 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 nominal redescriptions. For example, in (3)[3] 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.

(3) 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: Properties 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.

[2] In this respect our hypotheses are similar to those of Walker (1993) and Johnstone (1994), who hypothesized that speakers in task-oriented dialogues would restate information that was already in the common ground for the purpose of cuing particular task-related or conversational inferences.

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

Constraint Changes Hypothesis: The second type of task-related inference 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.[4] Our hypothesis is that this inference can also be cued by the content of a nominal expression when that expression realizes properties of a domain object that are not needed to identify which object is under discussion. For example, in (4) 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) and reduced the risk of the hearer missing the inference (Carletta, 1992).

(4) 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: Properties related to constraint changes are expressed in a context where the change is to be inferred by the hearer.

Commitment Hypothesis: 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 that a stage of the interaction is complete (Whittaker and Stenton, 1988; Jordan and Di Eugenio, 1997). Repeating properties 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 (5), 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).

(5) 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 commitment to a proposal, all the properties expressed in the proposal will be repeated.

Summarization Hypothesis: The second case in which a higher level subproblem was completed is illustrated by the summary in (6). 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.

(6) 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 properties for an item will be repeated.

Verification Hypothesis: 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; Walker, 1993). In the

[4] In (3) there is some explicit discussion about the color match goal.

COCONUT corpus, the hearer sometimes repeats the description in the turn immediately following. For example, in (5) 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 property information was correctly understood.

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

3 Experimental Approach

To verify our hypotheses about what could influence attribute selection, we undertook a two part corpus investigation. First, we did correlational studies on the corpus to get guidance on which of the hypotheses we should examine more closely. Our correlational studies showed that the contexts and attribute selections indicated in our hypotheses positively correlated for all but Verification and Summarization (Jordan, 2000a).

In the second part of our investigation, which is the subject of this paper, 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 could be similar 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. However, if the hypotheses were invalid then the resulting descriptions should only match the corpus as well as identificationally adequate descriptions that have some random attributes included. For example, if *my table* is identificationally adequate then it might also randomly include any of the remaining mutually known attributes as well (e.g. *my red table*, *my $250 table*).

We simulated selections for the COCONUT dialogues by using annotations about the discourse entities to be described and the contexts in which they appeared as input to the selection strategies we wished to test. We used existing annotations that were previously developed and tested for the COCONUT project as described in (Di Eugenio et al., 2000) as well as ones developed specifically for this research (Jordan, 2000b).[5]

One type of annotation feature we used to identify some of the contexts indicated in our hypotheses, were those that defined elements of the agreement process described in (Di Eugenio et al., 2000). First we will present high-level definitions of these agreement process elements and then we will explain how we used these definitions to identify the contexts.

- 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 context for the Summarization hypothesis is the most restrictive. An agreement must have been reached for an annotated action without the action being readdressed between the agreement and the Summarization. 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

[5] All of the annotations features were found to have good intercoder reliability.

be inferred and in the other the agreement is more explicit.

The Commitment context exists when a commitment is to be made and either 1) there is a previous proposal or unconditional commitment for the action involving the entity in the immediately previous turn and no other unrelated entities have been discussed for the action in the interim or 2) a speaker indicated unconditional commitment in his previous turn. This definition reflects commitment patterns described in (Di Eugenio et al., 2000).

The 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 et al., 2000), we identify proposals by looking for either a propose utterance, or an unconditional commitment utterance where the previous state for an annotated action is an unendorsed option or a partner decidable option. The alternatives are approximated by accumulating a list of the items evoked for each annotated action up until a propose or unconditional commitment occurs.

Once we have identified proposals and alternative solutions, next we check for contrasts to the alternative solutions and the partial solution. For color we compare the color of the proposed item to those items already selected for the room and the alternate items. If the proposed item matches the color of items already selected for the room while none of the alternates 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 alternate (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 alternate (i.e. the cheaper item may be preferred since it leaves some money for other purchases).

The remaining contexts are easier to recognize. The Verification context exists when an item was initially described in the immediately previous turn. The Domain constraint change context exists whenever an implicit constraint change is directly annotated.

We used the human generated descriptions in the COCONUT corpus to evaluate the descriptions created by the selection strategies we wished to test. 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 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 chose to include or exclude 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 description and N is the number of attributes that could be expressed for an entity. This response variable is called *match* in the experiments that follow. After doing an analysis of variance (Mat, 1998) on the results of experiments where we varied the selection strategy, we used multiple pairwise comparisons (MCA) (Hsu, 1996) [6] to locate where significant performance differences between strategies existed.[7] We display the results of the multiple comparisons as 95% confidence intervals, (e.g. as in Figure 3), which are always of the form:

> (estimate)±(critical point)×(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 this is indicated in each figure. [8]

[6] We used S-plus' multicomp function to perform the multiple comparisons (Mat, 1998).

[7] MCA is a standard statistical procedure for pairwise comparisons. It adjusts the ANOVA confidence intervals for error propagation (Hsu, 1996; Cohen, 1995).

[8] S-plus' multicomp function can optionally con-

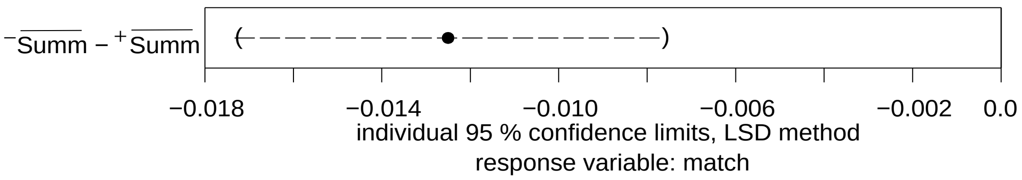

Figure 1: Difference between mean performances when excluding consideration of the Summarization hypothesis (-Summ) and including it (+Summ) for IINF

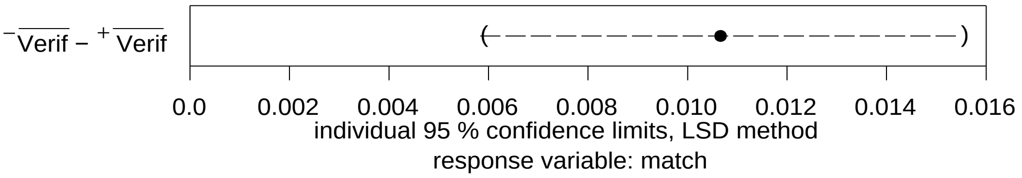

Figure 2: Difference between mean performances when excluding consideration of the Verification hypothesis (-Verif) and including it (+Verif) for IINF

Intervals in the figures that exclude zero indicate statistically significant performance differences. The labels on the y axis indicate the two levels or experimental factors for which the mean differences are shown. 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.

4 Simulation Results

First, we established a baseline where identification is the only goal. To address the identification goal, we used the incremental strategy of Dale and Reiter (1995) (**INC**). **INC** incrementally builds a description by checking an ordered list of attribute types and selecting 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.

Next, we created a parameterized selection strategy called *intentional influences* (**IINF**). **IINF** is 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. After selecting attributes as indicated by the included hypotheses, **IINF** then uses the **INC** strategy to determine if additional attributes are needed to rule out any remaining distractors.

To determine **IINF**'s parameter settings for this paper, we will accept the positive correlational results from the first part of our study and only skeptically test the negative ones. We found that Summarization had a clear positive influence while Verification had a clear negative one. For Summarization there is a significant difference in performance ($F = 25.71, p < .0000004$) and the performance comparison shown in Figure 1 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 2 indicates it is better not to consider Verification. As a result, the **IINF** strategy we test here will include all but the Verification hypothesis.

The final selection strategy that we will test is called randomized influences (**RINF**). It is motivated by our expectation that if the best combination of the communicative goals, as indirectly represented in our hypotheses, are not influential in selecting attributes then these additional goals would be the same as

sider all the valid methods to find the smallest critical point.

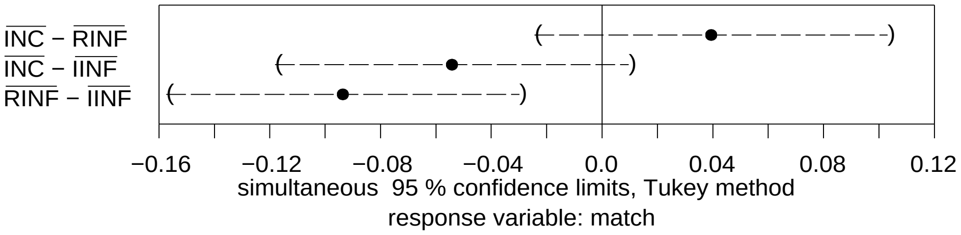

Figure 3: Differences between mean performances of the incremental model (INC), intentional influences model (IINF) and random influences model (RINF)

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

Table 1: Contributions of Goal Contexts to Redescriptions

IINF making random selections of the non-identificationally necessary attributes. To test this idea, the **RINF** strategy randomly decides whether to select a random number of attributes. As with **IINF**, it then uses **INC** to determine if additional attributes are needed to rule out distractors.

We found significant differences between the three selection strategies, **IINF**, **RINF**, and **INC** ($F = 6.05, p < .003$). As shown by the MCA confidence intervals in Figure 3, we found that **IINF** matched human descriptions significantly better than **RINF** whereas **INC** did not. **IINF**, while statistically similar to **INC**, also had a trend towards better matches when compared to **INC**.

Table 1 shows the relative contributions of the hypotheses included in **IINF** and the contribution of the identification goal within **IINF**. The contribution made by the identification goal includes both the cases in which identification was the only predicted goal (i.e. none of the contexts indicated in our four hypotheses applied for a particular description) and the cases in which additional attributes had to be added to ensure unique identifiability after the initial selections made in accordance with out hypotheses. Although the contributions made by the identification goal are smaller than one might expect, this does not mean that the identification goal was invalid for some redescriptions. Instead it indicates that the identification goal had been addressed already by some of the other applicable goals and reflects a type of potential economy that can be achieved when multiple goals influence one expression.

5 Conclusion

Our results indicate that we have identified a set of additional goals that can influence attribute selection for redescriptions. As we expected, allowing multiple goals to influence redescriptions did reflect allowable, alternative ways of identifying objects. In particular, we saw that the descriptions generated as a result of multiple goals and the descriptions generated to satisfy just the identification goal match equally well with what humans generate.

So far we have only partially addressed our original question about the relationship between multiple goals and nominal expressions. Among other things, we still need to ascertain the degree to which this relationship actually economizes the speaker's contribution and makes for more effective communication. In addition, we need to separate out the goals represented in the **IINF** selection strategy to see which cases are critical for ensuring an inference is made. However, to test these goals individually we need to collect more instances of redescriptions.

References

Douglas E. Appelt. 1985. *Planning English Sentences.* Cambridge University Press, Cambridge, U.K.

Susan E. Brennan. 1990. *Seeking and Providing Evidence for Mutual Understanding.* Ph.D. thesis, Stanford University Psychology Dept. Unpublished Manuscript.

Jean C. Carletta. 1992. *Risk Taking and Recovery in Task-Oriented Dialogue.* Ph.D. thesis, Edinburgh University.

Herbert H. Clark and Edward F. Schaefer. 1989. Contributing to discourse. *Cognitive Science*, 13:259–294.

Paul R. Cohen. 1995. *Empirical Methods for Artificial Intelligence.* MIT Press, Boston.

Robert Dale and Ehud Reiter. 1995. Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science*, 19(2):233–263, Apr–June.

Barbara Di Eugenio, Pamela W. Jordan, Johanna D. Moore, and Richmond H. Thomason. 1998. An empirical investigation of collaborative dialogues. In *ACL-COLING98, Proceedings of the Thirty-sixth Conference of the Association for Computational Linguistics*, Montreal, Canada, August.

Barbara Di Eugenio, Pamela W. Jordan, Richmond H. Thomason, and Johanna D. Moore. 2000. The agreement process: An empirical investigation of human-human computer-mediated collaborative dialogues. *To Appear in International Journal of Human-Computer Studies.*

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

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

Pamela W. Jordan and Barbara Di Eugenio. 1997. 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*, pages 81–84. The AAAI Press.

Pamela W. Jordan. 2000a. Influences on attribute selection in redescriptions: A corpus study. In *Proceedings of CogSci2000*, August.

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

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

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

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

Megan Moser and Johanna D. Moore. 1995. Investigating cue placement and selection in tutorial discourse. In *Proceedings of 33rd Annual Meeting of the Association for Computational Linguistics*, pages 130–135.

Rebecca J. Passonneau. 1995. Integrating Gricean and attentional constraints. In *Proceedings of IJCAI 95.*

Martha E. Pollack. 1991. Overloading intentions for efficient practical reasoning. *Noûs*, 25:513 – 536.

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

Marilyn A. Walker. 1992. Redundancy in collaborative dialogue. In *Fourteenth International Conference on Computational Linguistics*, pages 345–351.

Marilyn A. Walker. 1993. *Informational Redundancy and Resource Bounds in Dialogue.* Ph.D. thesis, University of Pennsylvania, December.

Steve Whittaker and Phil Stenton. 1988. Cues and control in expert client dialogues. In *Proc. 26th Annual Meeting of the ACL, Association of Computational Linguistics*, pages 123–130.