An Empirical Investigation of Proposals in Collaborative Dialogues

Barbara Di Eugenio **Pamela W. Jordan**
Johanna D. Moore **Richmond H. Thomason**
Learning Research & Development Center, and Intelligent Systems Program
University of Pittsburgh
Pittsburgh, PA 15260, USA
{dieugeni,jordan,jmoore,thomason}@isp.pitt.edu

Abstract

We describe a corpus-based investigation of proposals in dialogue. First, we describe our DRI compliant coding scheme and report our inter-coder reliability results. Next, we test several hypotheses about what constitutes a well-formed proposal.

1 Introduction

Our project's long-range goal (see http://www.isp.pitt.edu/~intgen/) is to create a unified architecture for collaborative discourse, accommodating both interpretation and generation. Our computational approach (Thomason and Hobbs, 1997) uses a form of weighted abduction as the reasoning mechanism (Hobbs et al., 1993) and modal operators to model context. In this paper, we describe the corpus study portion of our project, which is an integral part of our investigation into recognizing how conversational participants coordinate agreement. From our first annotation trials, we found that the recognition of "classical" speech acts (Austin, 1962; Searle, 1975) by coders is fairly reliable, while recognizing contextual relationships (e.g., whether an utterance accepts a proposal) is not as reliable. Thus, we explore other features that can help us recognize how participants coordinate agreement.

Our corpus study also provides a preliminary assessment of the Discourse Resource Initiative (DRI) tagging scheme. The DRI is an international "grassroots" effort that seeks to share corpora that have been tagged with the core features of interest to the discourse community. In order to use the core scheme, it is anticipated that each group will need to refine it for their particular purposes. A usable draft core scheme is now available for experimentation (see http://www.georgetown.edu/luperfoy/Discourse-Treebank/dri-home.html). Whereas several groups are working with the unadapted core DRI scheme (Core and Allen, 1997; Poesio and Traum, 1997), we have attempted to adapt it to our corpus and particular research questions.

First we describe our corpus, and the issue of tracking agreement. Next we describe our coding scheme and our intercoder reliability outcomes. Last we report our findings on tracking agreement.

2 Tracking Agreement

Our corpus consists of 24 computer-mediated dialogues[1] in which two participants collaborate on a simple task of buying furniture for the living and dining rooms of a house (a variant of the task in (Walker, 1993)). The participants' main goal is to negotiate purchases; the items of highest priority are a sofa for the living room and a table and four chairs for the dining room. The problem solving task is complicated by several secondary goals: 1) Match colors within a room, 2) Buy as much furniture as you can, 3) Spend all your money. A point system is used to motivate participants to try to achieve as many goals as possible. Each subject has a budget and inventory of furniture that lists the quantities, colors, and prices for each available item. By sharing this initially private information, the participants can combine budgets and select furniture from either's inventory. The problem is collaborative in that all decisions have to be consensual; funds are shared and purchasing decisions are joint.

In this context, we characterize an agreement as accepting a partner's suggestion to include a specific furniture item in the solution. In this paper we will focus on the issue of recognizing that a suggestion has been made (i.e. a proposal). The problem is not easy, since, as speech act theory points out (Austin, 1962; Searle, 1975), surface form is not a clear indicator of speaker intentions. Consider excerpt (1):[2]

(1) A: [35]: i have a blue sofa for 300.
[36]: it's my cheapest one.

B: [37]: I have 1 sofa for 350
[38]: that is yellow
[39]: which is my cheapest,
[40]: yours sounds good.

[35] is the first mention of a sofa in the conversa-

[1] Participants work in separate rooms and communicate via the computer interface. The interface prevents interruptions.

[2] We broke the dialogues into utterances, partly following the algorithm in (Passonneau, 1994).

tion and thus cannot count as a proposal to include it in the solution. The sofa A offers for consideration, is effectively proposed only after the exchange of information in [37]—[39].

However, if the dialogue had proceeded as below, [35'] would count as a proposal:

(2) B: [32']: I have 1 sofa for 350
[33']: that is yellow
[34']: which is my cheapest.

A: [35']: i have a blue sofa for 300.

Since context changes the interpretation of [35], our goal is to adequately characterize the context. For this, we look for guidance from corpus and domain features. Our working hypothesis is that for both participants context is partly determined by the domain reasoning situation. Specifically, if the suitable courses of action are highly limited, this will make an utterance more likely to be treated as a proposal; this correlation is supported by our corpus analysis, as we will discuss in Section 5.

3 Coding Scheme

We will present our coding scheme by first describing the core DRI scheme, followed by the adaptations for our corpus and research issues. For details about our scheme, see (Di Eugenio et al., 1997); for details about features we added to DRI, but that are not relevant for this paper, see (Di Eugenio et al., 1998).

3.1 The DRI Coding Scheme

The aspects of the core DRI scheme that apply to our corpus are a subset of the dimensions under *Forward-* and *Backward-Looking Functions*.

3.1.1 Forward-Looking Functions

This dimension characterizes the potential effect that an utterance U_i has on the subsequent dialogue, and roughly corresponds to the classical notion of an *illocutionary act* (Austin, 1962; Searle, 1975). As each U_i may simultaneously achieve multiple effects, it can be coded for three different aspects: *Statement, Influence-on-Hearer, Influence-on-Speaker*.

Statement. The primary purpose of *Statements* is to make claims about the world. *Statements* are subcategorized as an *Assert* when Speaker S is trying to change Hearer H's beliefs, and as a *Reassert* if the claim has already been made in the dialogue.

Influence-on-Hearer (I-on-H). A U_i tagged with this dimension influences H's future action. DRI distinguishes between S merely laying out options for H's future action (*Open-Option*), and S trying to get H to perform a certain action (see Figure 1). *Info-Request* includes all actions that request information, in both explicit and implicit forms. All other actions[3] are *Action-Directives*.

[3] Although this may cause future problems (Tuomela, 1995), DRI considers *joint* actions as decomposable into independent *Influence-on-Speaker / Hearer* dimensions.

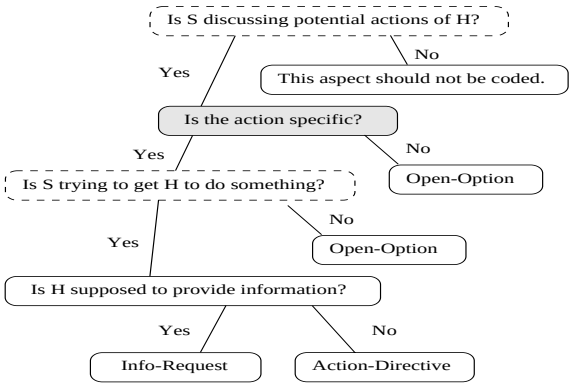

Figure 1: Decision Tree for Influence-on-Hearer

Influence-on-Speaker (I-on-S). A U_i tagged with this dimension potentially commits S (in varying degrees of strength) to some future course of action. The only distinction is whether the commitment is conditional on H's agreement (*Offer*) or not (*Commit*). With an *Offer*, S indicates willingness to commit to an action if H accepts it. *Commits* include promises and other weaker forms.

3.1.2 Backward Functions

This dimension indicates whether U_i is unsolicited, or responds to a previous U_j or segment.[4] The tags of interest for our corpus are:

- **Answer**: U_i answers a question.
- **Agreement**:

1. U_i *Accept/Reject*s if it indicates S's attitude towards a belief or proposal embodied in its antecedent.
2. U_i *Hold*s if it leaves the decision about the proposal embodied in its antecedent open pending further discussion.

3.2 Refinements to Core Features

The core DRI manual often does not operationalize the tests associated with the different dimensions, such as the two dashed nodes in Figure 1 (the shaded node is an addition that we discuss below). This resulted in strong disagreements regarding Forward Functions (but not Backward Functions) during our initial trials involving three coders.

Statement. In the current DRI manual, the test for *Statement* is whether U_i can be followed by "That's not true.". For our corpus, only syntactic imperatives or interrogatives were consistently filtered out by this purely semantic test. Thus, we refined it by appealing to syntax, semantics, and domain knowledge: U_i is a *Statement* if it is declarative

[4] Space constraints prevent discussion of segments.

and it is 1) past; or 2) non past, and contains a stative verb; or 3) non past, and contains a non-stative verb in which the implied action:
• does *not* require agreement in the domain;
• or is supplying agreement.
For example, *We could start in the living room* is not tagged as a statement if meant as a suggestion, i.e. if it requires agreement.

I-on-H and I-on-S. These two dimensions depend on the *potential action* underlying U_i (see the root node in Figure 1 for I-on-H). The initial disagreements with respect to these functions were due to the coders not being able to consistently identify such actions; thus, we provide a definition for actions in our domain,[5] and heuristics that correlate types of actions with I-on-H/I-on-S.

We have two types of potential actions: *put furniture item X in room Y* and *remove furniture item X from room Y*. We subcategorize them as *specific* and *general*. A *specific* action has all necessary parameters specified (*type, price* and *color* of item, and *room*). *General* actions arise because all necessary parameters are not set, as in *I have a blue sofa* uttered in a null context.

Heuristic for I-on-H (the shaded node in Figure 1). If H's potential action described by U_i is *specific*, U_i is tagged as *Action-Directive*, otherwise as *Open-Option*.

Heuristic for I-on-S. Only a U_i that describes S's *specific* actions is tagged with an *I-on-S* tag.

Finally, it is hard to offer comprehensive guidance for the test *is S trying to get H to do something?* in Figure 1, but some special cases can be isolated. For instance, when S refers to one action that the participants could undertake, but in the same turn makes it clear the action is not to be performed, then S is not trying to get H to do something. This happens in excerpt (1) in Section 2. A specific action (*get B's $350 yellow sofa*) underlies [38], which qualifies as an *Action-Directive* just like [35]. However, because of [40], it is clear that B is not trying to get A to use B's sofa. Thus, [38] is tagged as an *Open-Option*.

3.3 Coding for problem solving features

In order to investigate our working hypothesis about the relationship between context and limits on the courses of action, we coded each utterance for features of the problem space. Since we view the problem space as a set of constraint equations, we decided to code for the variables in these equations and the number of possible solutions given all the possible assignments of values to these variables.

The variables of interest for our corpus are the objects of type t in the goal to put an object in a room (e.g. var_{sofa}, var_{table} or var_{chairs}). For a solution to exist to the set of constraint equations, each var_i in the set of equations must have a solution. For example, if 5 instances of sofas are known for var_{sofa}, but every assignment of a value to var_{sofa} violates the budget constraint, then var_{sofa} and the constraint equations are unsolvable.

We characterize the solution size for the problem as determinate if there is one or more solutions and indeterminate otherwise. It is important to note that the set of possible values for each var_i is not known at the outset since this information must be exchanged during the interaction. If S supplies appropriate values for var_i but does not know what H has available for it then we say that no solution is possible at this time. It is also important to point out that during a dialogue, the solution size for a set of constraint equations may revert from determinate to indeterminate (e.g. when S asks what else H has available for a var_i).

[5] Our definition of actions does not apply to *Info-Requests*, as the latter are easy to recognize.

Stat.	I-on-H	I-on-S	Answer	Agr.
.83	.72	.72	.79	.54

Table 1: Kappas for Forward and Backward Functions

4 Analysis of the Coding Results

Two coders each coded 482 utterances with the adapted DRI features (44% of our corpus). Table 1 reports values for the Kappa (K) coefficient of agreement (Carletta, 1996) for Forward and Backward Functions.[6]

The columns in the tables read as follows: if utterance U_i has tag X, do coders agree on the subtag? For example, the possible set of values for *I-on-H* are: NIL (U_i is not tagged with this dimension), Action-Directive, Open-Option, and Info-Request. The last two columns probe the subtypes of Backward Functions: was U_i tagged as an answer to the same antecedent? was U_i tagged as *accept*ing, *reject*ing, or *hold*ing the same antecedent?[7]

K factors out chance agreement between coders; K=0 means agreement is not different from chance, and K=1 means perfect agreement. To assess the import of the values $0 < K < 1$ beyond K's statistical significance (all of our K values are significant at p=0.000005), the discourse processing community uses Krippendorf's scale (1980)[8], which dis-

[6] For problem solving features, K for two doubly coded dialogues was > .8. Since reliability was good and time was short, we used one coder for the remaining dialogues.

[7] In general, we consider 2 non-identical antecedents as equivalent if one is a subset of the other, e.g. if one is an utterance U_j and the other a segment containing U_j.

[8] More forgiving scales exist but have not yet been discussed by the discourse processing community, e.g. the one in (Rietveld and van Hout, 1993).

Stat.	I-on-H	I-on-S	Answer	Agr.
.68	. 71	N/S[a]	.81	.43

[a] N/S means not significant

Table 2: Kappas from (Core and Allen 97)

	Det	Indet	Unknown
AD+O	25	7	0
Open-Option	2	2	0
AD+C	10	2	0
Other	4	2	4

Table 3: Antecedents of Commit

counts any variable with K < .67, and allows tentative conclusions when .67 < K < .8 K, and definite conclusions when K≥.8. Using this scale, Table 1 suggests that Forward Functions and Answer can be recognized far more reliably than *Agreement*.

To assess the DRI effort, clearly more experiments are needed. However, we believe our results show that the goal of an adaptable core coding scheme is reasonable. We think we achieved good results on Forward Functions because, as the DRI enterprise intended, we adapted the high level definitions to our domain. However, we have not yet done so for Agreement since our initial trial codings did not reveal strong disagreements; now given our K results, refinement is clearly needed. Another possible contributing factor for the low K on Agreement is that these tags are much rarer than the Forward Function tags. The highest possible value for K may be smaller for low frequency tags (Grove et al., 1981).

Our assessment is supported by comparing our results to those of Core and Allen (1997) who used the unadapted DRI manual — see Table 2. Overall, our Forward Function results are better than theirs (the non significant K for I-on-S in Table 2 reveals problems with coding for that tag), while the Backward Function results are compatible. Finally, our assessment may only hold for task-oriented collaborative dialogues. One research group tried to use the DRI core scheme on free-flow conversations, and had to radically modify it in order to achieve reliable coding (Stolcke et al., 1998).

5 Tracking Propose and Commit

It appears we have reached an impasse; if human coders cannot reliably recognize when two participants achieve agreement, the prospect of automating this process is grim. Note that this calls into question analyses of agreements based on a single coder's tagging effort, e.g. (Walker, 1996). We think we can overcome this impasse by exploiting the reliability of Forward Functions. Intuitively, a U_i tagged as *Action-Directive + Offer* should correlate with a proposal — given that all actions in our domain are joint, an *Action-Directive* tag always co-occurs with either *Offer (AD+O)* or *Commit (AD+C)*. Further, analyzing the antecedents of *Commit*s should shed light on what was treated as a proposal in the dialogue. Clearly, we cannot just analyze the antecedents of *Commit* to characterize proposals, as a proposal may be discarded for an alternative.

To complete our intuitive characterization of a proposal, we will assume that for a U_i to count as a well-formed proposal (WFP), the context must be such that enough information has already been exchanged for a decision to be made. The feature *solution size* represents such a context. Thus our first testable characterization of a WFP is:

1.1 U_i counts as a WFP if it is tagged as *Action-Directive + Offer* and if the associated solution size is determinate.

To gain some evidence in support of 1.1, we checked whether the hypothesized WFPs appear as antecedents of *Commits*.[9] Of the 32 *AD+O*s in Table 3, 25 have determinate solution size; thus, WFPs are the largest class among the antecedents of *Commit*, even if they only account for 43% of such antecedents. Another indirect source of evidence for hypothesis 1.1 arises by exploring the following questions: are there any WFPs that are not committed to? if yes, how are they dealt with in the dialogue? If hypothesis 1.1 is correct, then we expect that each such U_i should be responded to in some fashion. In a collaborative setting such as ours, a partner cannot just ignore a WFP as if it had not occurred. We found that there are 15 *AD+O*s with determinate solution size in our data that are not committed to. On closer inspection, it turns out that 9 out of these 15 are actually indirectly committed to. Of the remaining 6, four are responded to with a counterproposal (another *AD+O* with determinate solution size). Thus only two are not responded to in any fashion. Given that these 2 occur in a dialogue where the participants have a distinctively non-collaborative style, it appears hypothesis 1.1 is supported.

Going back to the antecedents of *Commit* (Table 3), let's now consider the 7 indeterminate *AD+O*s. They can be considered as *tentative* proposals that need to be negotiated.[10] To further refine our characterization of proposals, we explore the hypothesis:

[9] Antecedents of *Commits* are not tagged. We reconstructed them from either *variable* tags or when U_i has both Commit and Accept tags, the antecedent of the Accept.

[10] Because of our heuristics of tagging specific actions as *ActionDirectives*, these utterances are not *Open-Options*.

1.2 When the antecedent of a *Commit* is an *AD+O* and indeterminate, the intervening dialogue renders the solution size determinate.

In 6 out of the 7 indeterminate antecedent *AD+O*s, our hypothesis is verified (see excerpt (1), where [35] is an *AD+O* with indeterminate solution size, and the antecedent to the *Commit* in [40]).

As for the other antecedents of *Commit* in Table 3, it is not surprising that only 4 *Open-Options* occur given the circumstances in which this tag is used (see Figure 1). These *Open-Options* appear to function as *tentative* proposals like indeterminate *AD+O*s, as the dialogue between the *Open-Option* and the *Commit* develops according to hypothesis 1.2. We were instead surprised that *AD+C*s are a very common category among the antecedents of *Commit* (20%); the second commit appears to simply reconfirm the commitment expressed by the first (Walker, 1993; Walker, 1996), and does not appear to count as a proposal. Finally, the *Other* column is a collection of miscellaneous antecedents, such as *Info-Requests* and cases where the antecedent is unclear, that need further analysis. For further details, see (Di Eugenio et al., 1998).

6 Future Work

Future work includes, first, further exploring the factors and hypotheses discussed in Section 5. We characterized WFPs as *AD+O*s with determinate solution size: a study of the features of the dialogue preceding the WFP will highlight how different options are introduced and negotiated. Second, whereas our coders were able to reliably identify Forward Functions, we do not expect computers to be able to do so as reliably, mainly because humans are able to take into account the full previous context. Thus, we are interested in finding correlations between Forward Functions and "simpler" tags.

Acknowledgements

This material is based on work supported by the National Science Foundation under Grant No. IRI–9314961. We wish to thank Liina Pyllkänen for her contributions to the coding effort, and past and present project members Megan Moser and Jerry Hobbs.

References

John L. Austin. 1962. *How to Do Things With Words*. Oxford University Press, Oxford.

Jean Carletta. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2).

Mark G. Core and James Allen. 1997. Coding dialogues with the DAMSL annotation scheme. AAAI Fall Symposium on *Communicative Actions in Human and Machines*, Cambridge MA.

Barbara Di Eugenio, Pamela W. Jordan, and Liina Pylkkänen. 1997. The COCONUT project: Dialogue annotation manual. http://www.isp.pitt.edu/~intgen/research-papers.

Barbara Di Eugenio, Pamela W. Jordan, Richmond H. Thomason, and Johanna D. Moore. 1998. The Acceptance cycle: An empirical investigation of human-human collaborative dialogues. Submitted for publication.

William M. Grove, Nancy C. Andreasen, Patricia McDonald-Scott, Martin B. Keller, and Robert W. Shapiro. 1981. Reliability studies of psychiatric diagnosis. theory and practice. *Archives General Psychiatry*, 38:408–413.

Jerry Hobbs, Mark Stickel, Douglas Appelt, and Paul Martin. 1993. Interpretation as abduction. *Artificial Intelligence*, 63(1–2):69–142.

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

Rebecca J. Passonneau. 1994. Protocol for coding discourse referential noun phrases and their antecedents. Technical report, Columbia University.

Massimo Poesio and David Traum. 1997. Representing conversation acts in a unified semantic/pragmatic framework. AAAI Fall Symposium on *Communicative Actions in Human and Machines*, Cambridge MA.

T. Rietveld and R. van Hout. 1993. *Statistical Techniques for the Study of Language and Language Behaviour*. Mouton de Gruyter.

John R. Searle. 1975. Indirect Speech Acts. In P. Cole and J.L. Morgan, editors, *Syntax and Semantics 3. Speech Acts*. Academic Press.

A. Stolcke, E. Shriberg, R. Bates, N. Coccaro, D. Jurafsky, R. Martin, M. Meteer, K. Ries, P. Taylor, and C. Van Ess-Dykema. 1998. Dialog act modeling for conversational speech. AAAI Spring Symposium on *Applying Machine Learning to Discourse Processing*.

Richmond H. Thomason and Jerry R. Hobbs. 1997. Interrelating interpretation and generation in an abductive framework. AAAI Fall Symposium on *Communicative Actions in Human and Machines*, Cambridge MA.

Raimo Tuomela. 1995. *The Importance of Us*. Stanford University Press.

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

Marilyn A. Walker. 1996. Inferring acceptance and rejection in dialogue by default rules of inference. *Language and Speech*, 39(2).