A Two-level Approach to Coding Dialogue for Discourse Structure: Activities of the 1998 DRI Working Group on Higher-level Structures*

David R. Traum
University of Maryland
traum@cs.umd.edu

Christine H. Nakatani
Bell Laboratories, Lucent Technologies
chn@research.bell-labs.com

Abstract

This paper presents a novel two-level scheme for coding discourse structure in dialogue, which was created by the authors for the discourse structure subgroup of the 1998 DRI meeting on dialogue tagging. We discuss the theoretical motivations and framework for the coding proposal, and then review the results of coding exercises performed by the 1998 DRI discourse structure subgroup using the new manual. Finally, we provide suggestions for improving the scheme arising from the working group activities at the third DRI meeting.

1 Introduction

A two-level scheme for coding discourse structure in dialogue has been proposed and undergone initial testing within the DRI effort. In particular, the higher-level structures working group of the third DRI was charged with the task of creating a coding scheme concerned exclusively with the discourse structure of *dialogue*. Finding a good starting point for a consensus coding scheme for discourse structure in dialogue was a non-trivial task. Most discourse structure schemes in fact were geared toward monologue, and most dialogue coding schemes omitted the higher-level structures that were essential to the monologue schemes, or provided only genre or domain-specific higher-level structures.

Given the limited amount of work in this area, it was impossible to attempt a comprehensive coding scheme for all aspects of discourse structure in dialogue. Instead, we were guided by an analysis of what choices needed to be made in creating a coding scheme. (Traum, 1998) identifies three dimensions along which discourse structure schemes can be classified: *granularity, content, structuring mechanisms.*

- *Granularity:* how much material (time, text, turns, etc.) is covered by the units (minimum, maximum, and average)? Granularity ranges were divided roughly into three categories:

  Micro - roughly within a single turn

  Meso - roughly an exchange, IR-unit, "game", or short "sub-dialogue",

  Macro - coherent larger spans, related to overall dialogue purposes.

- *Content:* what is this a structure *of* (e.g., intentions, accessibility, effects, etc.)?
- *Structuring mechanisms:* What kinds of units and structuring principles are used (e.g., flat, set inclusion, hierarchical/CFG structuring, relational)? How many primitive types of units are allowed (one basic unit type, two or three types of units, or several types)?

This multi-dimensional space was then used to classify different extant coding schemes as to which aspects they are concerned with.

Guided by this principled survey of various schemes, we decided on an objective of defining a pair of coupled schemes at the meso- and macro-levels in order to create a dialogue-oriented scheme for discourse structure analysis. We felt the micro-level of analysis was addressed by the dialogue acts coding effort of DRI, and it seemed most productive to build meso- and macro-levels on top of that, in an independent manner, to see what synergy might arise. It did not seem most fruitful to code the same

*The discourse structure working group was chaired by Christine Nakatani (Bell Laboratories, Lucent Technologies) and co-chaired by David Traum (U Maryland). Pre-meeting group participants also included Jean Carletta (U Edinburgh), Jennifer Chu-Carroll (Bell Laboratories, Lucent Technologies), Peter Heeman (Oregon Graduate Institute), Julia Hirschberg (AT&T Labs), Masato Ishizaki (JAIST), Diane Litman (AT&T Labs), Owen Rambow (Cogentex), Jennifer Venditti (Ohio State U), Marilyn Walker (At&T Labs), Gregory Ward (Northwestern U). Participants at the meeting also included Ellen Bard (U Edinburgh), Yasuo Horiuchi (Chiba U), Koichi Hoshida (ATR), Yasuhiro Katagisi (NTT), Kikuo Maekawa (NLRI), Michael Strube (U Pennsylvania), Masafumi Tamato (NTT), Yuki Tateishi (Tokyo U), and Takahiro Wakao (TAO).

content at three different levels, or to code three types of content at the macro-level without making any attempt to relate that coding to other schemes in development within the DRI initiative.

Thus, for our starting point we proposed two original coding schemes within this multi-dimensional space. One scheme which has as content *Grounding* (Clark and Schaefer, 1989; Traum, 1994), operated at a *meso* level of granularity, and used non-hierarchical (and possibly discontinuous) utterance sets as its structuring principle. The second scheme concerned *intentional/informational structure* (Grosz and Sidner, 1986; Nakatani et al., 1995) as content, operated at a *macro* level of granularity, and was structured as hierarchical trees (with annotations for capturing discontinuities). In addition, these two schemes were linked by using the resulting structures from meso-level analysis as basic input for macro-level analysis.

There were several factors motivating the decision to use these particular facets of discourse structure for initial analysis. First, considering intentions, it is clear that aspects of dialogue at all levels of granularity relate to the intentions of the participants. However, not all of these intentional aspects are attuned to well-behaved plan-like structures. One issue is whose intention is under consideration: the speaker, the hearer, or the collaborative "team" of the speaker and hearer together. It is only at the level of grounded content that some sort of joint or shared intentional structure is really applicable. Below this level, one may only properly talk of individual intentions, even though those intentions may be subservient to joint goals (or goals of achieving sharedness). Thus taking grounded units (achieved at the meso-range) as a starting point for the coding of intentional structure is a natural basis for the study of joint intentional structure. Individual intentions at a lower level, especially those relating to communication management rather than task are expected to be captured within the dialogue act level of the DRI coding scheme (Discourse Resource Initiative, 1997; Allen and Core, Draft 1997). Likewise, the phenomena of grounding can occur on multiple levels. However, since macro-level phenomena (such as task summarization) differ from more local feedback phenomena (including acknowledgments and repairs), restricting the grounding-relating coding to the meso-level allows for a more tractable effort.

While examining intentional structure at the macro range and grounding structure at a meso range thus had independent motivations, the coding scheme used for this subgroup was designed to test a further novel and previously untested hypothesis that the units of achieving common ground would serve as an appropriate type of basic unit for intentional analysis. Since the phenomena of grounding and intentional task-related structure are somewhat independent, there is reason to believe the structures might not align properly. However, given the utility of having an appropriate meso-level starting point for intentional structure, and lacking any compelling counter-examples, we decided to put the hypothesis to the test in the coding exercises.

2 The coding scheme

The coding scheme used for pre-meeting coding exercises is defined in (Nakatani and Traum, 1999), which was distributed to the group members prior to coding assignments. As mentioned above, this included two levels of coding, common ground units (**CGUs**) at the meso-level, and intentional/informational units (**IUs**) at the macro-level.

Here we provide a brief summary of these coding schemes. Interested parties are referred to the manual (Nakatani and Traum, 1999) for detailed instructions and examples. There are three stages of coding, which must be performed in sequence. First, a preparatory *tokenization* phase, in which the dialogue is segmented into speaker turns and utterance tokens within the turns, each token being given a label. This was used as input for the coding of CGUs, in which utterance tokens were gathered together in units of tokens which together served to add some material to the common ground. Finally, the results of CGU coding was used as input for IU Coding, in which hierarchical intentional structure was built from either CGUs or smaller IUs. Each of these processes is briefly described in the subsections below.

2.1 Common Ground Units (CGUs)

A Common Ground Unit (CGU) contains all and only the utterance tokens needed to *ground* (that is, make part of the common ground) some bit of content. This content will include the initial token of the unit, plus whatever additional content is added by subsequent tokens in the unit and added to the common ground at the same time as the initiating token. The main coherence principle for CGUs is thus not directly related to the coherence of the content itself (this kind of coherence is handled at the micro and macro levels), but whether the content is added to the common ground in the same manner (e.g., with the same acknowledgment utterance).

CGUs will require at least some initiating material by one conversational participant (the initiator), presenting the new content, as well as generally some *feedback* (Allwood et al., 1992), or acknowledgment, by the other participant.

The following principles in (1) summarize the decision procedures for how to code an utterance token with respect to existing or new CGUs:

(1) 1. **If** the token contains *new* content, and there is no accessible ungrounded CGU, the contents of which could be acknowledged together with the current token

then create a new CGU, and add this token to it.

2. **if** there is an accessible CGU for which the current token:
   (a) acknowledges the content
   (b) repairs the content
   (c) cancels the CGU (in this case, also put a * before the CGU marker, to indicate that it is canceled).
   (d) continues the content, in such a fashion that all content could be grounded together (with the same acknowledgment)

   then add this token to the CGU
   otherwise, do not add this token to the CGU

Note that these rules are not mutually exclusive: more than one may apply, so that a token can be added to more than one CGU.

CGUs are similar in many respects to other meso-level coding schemes, such as *initiative-response* in the LINDA coding scheme (Ahrenberg et al., 1990; Dahlbäck and Jönsson, 1998), or *conversational games* (Carletta et al., 1997). However, there are some important differences. In terms of content, CGUs cover only grounding, while the LINDA scheme covers initiative more generally, and the HCRC game structure codes achievement of dialogue purposes. Several authors (e.g., (Allwood et al., 1992; Clark, 1994; Dillenbourg et al., 1996), consider multiple levels of coordination in dialogue, including roughly those of *contact, perception, understanding, and attitudinal reaction*. Grounding (which is what CGUs capture) is mainly concerned with the understanding level (and also the perception of messages), while there is a large part of the notion of *response* that is concerned with attitudinal reaction and not strictly mutual understanding.

There are also differences in the *structuring mechanisms* used. In the LINDA coding scheme, IR units consist of trees, which may contain embedded IR units as constituents. The HCRC scheme does not require a strict tree structure, but also allows embedded games, when one game is seen as subordinate to the main purpose of another. In contrast, CGUs are "flat" structures, consisting only of a set of utterances which work together to add some material to common ground. Moreover, a single utterance can be part of multiple (non-nested) CGUs. For example, except for very short reactions which are expressed in the same locution with the feedback signal of understanding, the grounding of the reaction itself will also constitute a separate CGU. More concretely, consider a suggestion followed by a refinement by another speaker. The refinement indicates understanding of the original, and is thus part of the prior CGU, which presents the original, but it also introduces new material (the refinement itself), and thus also initiates a new CGU, which requires further signals of understanding to be added to the common ground.

Both of these differences in content and structuring mechanisms can lead to differences in the kinds of units that would be coded for a given dialogue fragment. For example, a question/answer/followup sequence might be one IR-unit or game but two CGUs (one to ground the question, and one to ground the answer). Likewise, a unit including a repair might be coded as two (embedded) IR-units or games, but only a single CGU.

It remains an open question as to whether CGUs or one of these other meso-level units might be the most appropriate building block for macro-level intentional structure. One reason to think that CGUs might be more appropriate, though, is the use of non-hierarchical units, which avoids the question of *which* level of unit to use as starting point.

2.2 Intentional/Informational Units (IUs)

Macro-level of discourse structure coding involves reasoning about the relationships amongst the pieces of information that have been established as common ground. This is achieved by performing a *topic-structure* or *planning-based* analysis of the content of the CGUs, to produce a hierarchy of CGUs in a well-formed tree data structure. Such analysis proceeds in similar fashion to the intention-based methodology outlined in (Nakatani et al., 1995), but there are some crucial differences. The coding scheme of (Nakatani et al., 1995) was developed for monologic discourse, and is not directly applicable to dialogue. In particular, there is the general problem in dialogue of associating the individual intentions of the participants with the overall structure. We use CGUs as a starting point helps establish the relevant intentions as a kind of joint intentional structure. While CGU analysis concentrates on establishing *what* is being said at the level of information exchange, macro-level analysis goes beyond this to establish relationships at a higher-level, namely relationships amongst CGUs (instead of utterance-tokens) and relationships amongst groups of CGUs. These relationships may be both informational and intentional. Thus, we refer to groupings of CGUs at the lowest level of macro-structure as I-UNITS (IUs), where "I" stands for either informational or intentional.

IU trees are created by identifying certain kinds of discourse relations. Following (Grosz and Sidner, 1986), macro-level analysis captures two fundamental intentional relations between I-units, those of *domination* (or parent-child) and *satisfaction-precedence* (or sibling) relations. The corresponding informational relations are *generates* and *enables* (Pollack, 1986; Goldman, 1970). More concretely, the domination relation can be elaborated in a planning-based framework as holding between a *subsidiary* plan and its parent, in which the completion of one plan contributes to the completion of its parent plan; the satisfaction-precedence relation can be elaborated as the temporal dependency between two plans (Lochbaum, 1994). As is often the case, when a temporal dependency cannot be strictly established, two IUs will be placed in a sibling relationship by virtue of their each being in a subsidiary relationship with the same dominating IU.

I-unit analysis consists of identifying the higher-level intentional/informational structure of the dialogue, where each I-unit (IU) in the macro structure achieves a joint (sub)goal or conveys information necessary to achieve a joint (sub)goal. The following schema captures the decision process for IU coding:

- Establish problem to be collaboratively solved, or *joint goal*.
- Negotiate how to achieve joint goal.
  This may involve:
  1. Deciding which (of possibly several) recipe(s) for action to use,
  2. Deciding how to implement a recipe in the participants' domain by instantiating or identifying constraints and parameters of the recipe (e.g. deciding which of two engines to move to the orange warehouse),
  3. Breaking the plan down into subplans, whose own achievements can be similarly negotiated at the subtask level.
- Confirm achievement of (or failure to achieve) joint goal.

This schema explicitly accommodates the inferential interface between the intentional and informational levels of analysis. For example, intentional and informational relations blend as siblings at the level of choosing and implementing a recipe and breaking down a plan into subplans. This reflects the simple fact that achieving a goal via action requires knowledge of the world (e.g. identification of objects), knowledge of how to act in the world (i.e. knowledge of recipes), and knowledge of how to reason about complex relations among actions (i.e. the ability to plan and re-plan). In sum, the blending of intentional and informational relations in IU coding is an original theoretical aspect of this coding scheme.

3 Coding exercises

In order to familiarize the group members with the coding schemes and provide some initial data for discussion, several coding exercises were performed, divided into two sets of two dialogues each – first TOOT and TRAINS, second Verbmobil (IU on common provided CGUs) and Maptask (only a fragment, no IU coding). These dialogues are all roughly characterizable as "task-oriented", although the tasks are quite varied.

The TRAINS dialogue was taken from the TRAINS-93 Corpus by the University of Rochester (Heeman and Allen, 1994; Heeman and Allen, 1995). TRAINS dialogs deal with tasks involving manufacturing and shipping goods in a railroad freight system. TRAINS dialogs consist of two human speakers, the system and the user. The user is given a problem to solve and a map of the world. The system is given a more detailed map and acts as a planning assistant to the user. Additional online information about the dialogues can be found at
`http://www.cs.rochester.edu/research/speech/93dialogs/` and about the trains project as a whole at
`http://www.cs.rochester.edu/research/trains/`

Toot dialogues are Human-Computer spoken dialogues, in which the computer system (S) finds Amtrak rail schedules via internet, according to specifications provided by the human user (U). The Toot system is described in (Litman et al., 1998). The dialogue we used for coding, was provided by Diane Litman of AT&T Research.

The Verbmobil project is a long term effort to develop a mobile translation system for spontaneous speech in face-to-face situations. The current domain of focus is scheduling business meetings. To support this goal, some English human-human dialogs were collected in this domain. More information about the Verbmobil project can be found online at `http://www.dfki.uni-sb.de/verbmobil/`. In the dialogue we coded, the two speakers try to establish a time and place for a meeting.

The DCIEM Map Task dialogs from which the one we coded (d204), was drawn were collected in Canada and consist of pairs of Canadian army reservists collaborating to solve a problem. Both reservists have a map but the maps are not identical in terms of the landmarks present. One participant is designated the direction giver, G and has a path marked on his map. The goal is for the other participant, the direction follower, F to trace this route on his map even though he can only communicate with G via speech; i.e., these are not face to face conversations. Only the opening portion of the dialogue was coded, due to the length. More information about the DCIEM Map Task corpus can be found online at `http://www.hcrc.ed.ac.uk/Site/MAPTASKD.html`.

A fragment taken from the Verbmobil Dialogue, along with CGU and IU coding for this fragment is shown in Figure 1. Note that some utterances (e.g., A.11.1) appear in multiple cgus (serving an acknowledgment function for one and a proposal function for the other), and some utterances (e.g., B.12.2) do not appear in any.

	TRAINS			TOOT			Verbmobil			Maptask		
	B	M	E	B	M	E	B	M	E	B	M	E
PA	0.83	0.87	0.85	0.79	0.81	0.78	0.79	0.78	0.89	0.69	0.74	0.79
PE	0.50	0.65	0.51	0.50	0.52	0.50	0.57	0.51	0.58	0.54	0.52	0.56
κ	0.66	0.62	0.69	0.58	0.60	0.56	0.52	0.56	0.74	0.34	0.45	0.52

Table 1: CGU Inter-coder Reliability

Verbmobil Dialogue r148c

```
...
A.9.2  want to have lunch
B.10.1  that sounds pretty good
B.10.2  are you available just before noon
A.11.1  we can meet at noon
B.12.1  sounds good
B.12.2  uhh
B.12.3  on campus or off
A.13.1  your choice
...
```

CGU and IU coding

```
iu.1    "plan to meet (again)"
  ...
  iu.1.2  "set meeting time"
  cgu7       A.9.2, B.10.1 "suggest lunch"
  cgu8       B.10.2, A.11.1 "suggest time"
  cgu9       A.11.1, B.12.1 "meet at noon"
  iu.1.3  "select place for lunch"
  cgu10      B.12.3, A.13.1  "on campus?"
  ...
```

Figure 1: Verbmobil CGU and IU coding

3.1 CGU Coding Analysis

The inter-coder reliability of CGU coding was quite variable between the different dialogues and for different stretches within some of the dialogues. Results ranged from segments in which all coders coded identically to a few segments (for Maptask and Toot) in which all coders coded some aspect differently. This section outlines some of the qualitative and quantitative analysis done on the CGU coding for the four dialogues presented in the previous section.

3.1.1 Inter-coder Reliability

It was a bit challenging to devise a meaningful measure of inter-coder reliability for the CGU coding task. While it is simple to count how many coders chose to include a particular unit, there is no easy way to devise an expected agreement for such a unit. Table 2 shows the average ratio of coders per CGU coded by any of the coders. It is not clear how to interpret this number, however, since if a particular unit was included only by a small amount of coders, that means that there was fairly high agreement among the other coders *not* to include it.

Simply marking down boundary points of units would also not work well, since CGUs are allowed to be both overlapping and discontinuous. Instead, a *pseudo-grounding acts* scheme was induced, considering whether an utterance token *begins*, *continues* or *completes* a CGU. This is fueled by the observation that, while a token could appear in multiple CGUs, it doesn't generally perform the same function in each of them. This is not explicitly ruled out but does seem to be the case, perhaps with one or two exceptions. So, each token is scored as to whether or not it appeared (1) as the first token in a CGU (2) as the last token in a CGU and/or (3) in a CGU in neither the first or last position.

This system seems sufficient to count as the same all identified CGUs that are the same, and to assess penalties for all codings that differ, though it is not clear that the weighting of penalties is necessarily optimal (e.g., leaving out a *middle* counts only one point of disagreement, but leaving out an *end* counts as two, since the next to last, gets counted as an *end* rather than a *middle*).

From this, it was possible to compute agreement and expected agreement (by examining the relative frequencies of these tags), and thus Kappa (Siegel and Castellan, 1988). The numbers for the group as a whole are shown in table 1 Systematic individual pairwise agreement or cluster analysis was not performed, however some of the pairwise numbers are above 0.8 for some dialogues.

From this table it is clear that the ending points of CGUs in verbmobil has fairly high agreement, as does the TRAINS dialogue overall, whereas Maptask has fairly low agreement, especially for CGU beginnings.

3.2 IU Coding Analysis

IU analysis was carried out on the Toot, Trains and Verbmobil dialogues. However, as noted, only the IU analysis on Verbmobil was conducted starting with

Dialogue	avg %
TRAINS	0.41
TOOT	0.36
Verbmobil	0.30
MAPTASK	0.26

Table 2: Average coders per proposed CGU

uniform IUs for all the coders. Thus, the reliability for IU coding could be quantitatively measured for the Verbmobil dialogue only. Nine coders provided IU trees starting from identical CGUs.

Following the methodology in (Hirschberg and Nakatani, 1996), we measured the reliability of coding for a linearized version of the IU tree, by calculating the reliability of coding of IU beginnings using the kappa metric. We calculated the observed pairwise agreement of CGUs marked as the beginnings of IUs, and factored out the expected agreement estimated from the actual data, giving the pairwise kappa score.

Table 3 gives the raw data on coders marking of IU beginnings. For each CGU, a "1" indicates that it was marked as an IU-initial CGU by a given coder. A "0" indicates that it was not marked as IU-initial.

	Coder									
CGU	1	2	3	4	5	6	7	8	9	TOTAL
1:	1	1	1	1	1	1	1	1	1	9/9
2:	0	0	0	0	0	0	0	0	0	0/9
3:	0	1	0	0	1	0	0	0	1	3/9
4:	0	0	0	0	0	0	0	0	0	0/9
5:	0	1	0	0	1	0	0	0	1	3/9
6:	0	0	0	0	0	0	0	0	0	0/9
7:	1	1	1	1	0	1	1	1	1	8/9
8:	1	1	0	0	0	0	1	0	0	3/9
9:	0	0	1	0	0	0	0	0	0	1/9
10:	1	1	1	1	1	1	1	1	1	9/9
11:	0	0	0	0	0	0	0	0	0	0/9
12:	0	0	1	0	0	0	0	0	1	2/9
13:	0	1	0	0	0	0	0	0	0	1/9
14:	1	1	0	1	0	1	1	1	1	7/9
15:	1	1	1	1	1	1	1	1	1	9/9

Table 3: Summary of IU coding for all coders (1=IU-initial, 0=non-IU-initial)

Table 4 shows the figures on observed pairwise agreement, or the percentage of the time both coders agreed on the assignment of CGUs to IU-initial position.

We calculated the expected probability of agreement for IU-initial CGUs to be P(E)=.375, based on the actual Verbmobil codings. Given P(E), kappa

ID	1	2	3	4	5	6	7	8	9
1	1	.8	.73	.93	.6	.93	.93	.93	.73
2		1	.53	.73	.67	.73	.73	.73	.8
3			1	.8	.6	.8	.67	.8	.73
4				1	.67	1	.87	1	.8
5					1	.67	.67	.67	.73
6						1	.87	1	.8
7							1	.87	.67
8								1	.8
9									1

Table 4: Observed agreement for IU-initial CGUs

ID	1	2	3	4	5	6	7	8	9
1	1	.7	.57	.89	.36	.89	.89	.89	.57
2		1	.25	.57	.47	.57	.57	.57	.68
3			1	.68	.36	.68	.47	.68	.57
4				1	.47	1	.79	1	.68
5					1	.47	.47	.47	.57
6						1	.79	1	.68
7							1	.79	.47
8								1	.68
9									1

Table 5: Pairwise kappa scores

scores can be computed. Table 5 shows the kappa scores measuring the reliability of the codings for each pair of labelers.

As the kappa scores show, there is some individual variation in IU coding reliability. On average, however, the kappa score for pairwise coding on IU-initial CGUs is .64, which is moderately reliable but shows room for improvement.

By examining Table 3, it can be seen that there was in fact always a decisive majority label for each CGU, i.e. there are no CGUs on which the coders were split into two groups of four and five in their coding decision for IU-initial CGUs. A weaker reliability metric on the pooled data from nine coders, therefore, would provide a reliable *majority* coding on this dialogue (see (Passonneau and Litman, 1997) for discussion of how reliability is computed for pooled coding data). In fact, for the group of six coders who showed the most inter-coder agreement, the average pairwise kappa score is .80, which is highly reliable.

4 Summary and Future Work

In addition to the quantitative analysis of codings, the subgroup at the 1998 DRI meeting reiterated some goals for the scheme in general and made progress on several open theoretical issues.[1] First and foremost, it was agreed upon that CGU analysis at the meso-level allowed coders to abstract the "messy" bits of dialogue (e.g., local repair, turn-taking, grounding) into common ground units, making the structures at both the meso- and macro-levels cleaner. The consensus was that many NLP applications would benefit from this abstraction, which can help separate to a large degree the processing of dialogic phenomena from the processing of intentions and informational units at the dialogue planning level.

As for theoretical issues, the subgroup laid out initial proposals for exploring the interface between Damsl tagging at the dialogue act micro-level, and

[1] Full details of the subgroup proceedings can be found in the DRI report of the 1998 meeting, also available from the first author.

CGU analysis at the meso-level. One important open issue was whether to modify the coding scheme to identify different types of acknowledgments separately, especially when the acknowledgment function was parasitic on a more direct relation, such as an answer to a question. It was found that alternative proposals for placing CGU boundaries patterned with differences in backward- and forward-looking properties of the ambiguous tokens. The general principle that was agreed upon was that we should investigate further the situations in which dialogue act coding can serve as the basis for CGU coding decisions, just as CGU codings serve as the primitive units for constraining IU analysis in a substantial way. A more general principle was to identify when independent decisions at one level could influence the coding decisions at a second level, e.g. when an IU boundary resolved a difficult CGU boundary decision. Defining non-circular coding guidelines appears feasible, if difficult.

While the reliability results presented here are already close to acceptable, directions for future work are clear. In particular, extensions to include additional dimensions of dialogue content would be desirable; the current scheme considers only grounding at the meso-range, and information/intention content at the macro-range. Secondly, we expect refinement and revision of the initial coding manual, (Nakatani and Traum, 1999), will facilitate both greater reliability and utility of the two levels we do cover. We hope other researchers will explore whether a more productive synergy can be found between the two levels, both in theory and in practice. The relation we hypothesize between the two levels, and our supposition that important relations may be found between micro-level schemes and the two-level scheme posited here, lay the groundwork for more focused investigations of coding schemes for discourse structure in dialogue than have previously existed within the DRI initiative.

References

Lars Ahrenberg, Nils Dahlbäck, and Arne Jönsson. 1990. Discourse representation and discourse management for a natural language dialogue system. In *Proceedings of the Second Nordic Conference on Text Comprehension in Man and Machine.*

James Allen and Mark Core. Draft, 1997. Draft of damsl: Dialog act markup in several layers. available through the WWW at: http://www.cs.rochester.edu/research/trains/annotation.

Jens Allwood, Joakim Nivre, and Elisabeth Ahlsen. 1992. On the semantics and pragmatics of linguistic feedback. *Journal of Semantics*, 9.

Jean Carletta, Amy Isard, Stephen Isard, Jacqueline C. Kowtko, Gwynewth Doherty-Sneddon, and Anne H. Anderson. 1997. The reliability of a dialogue structure coding scheme. *Computational Linguistics*, 23(1):13–31.

Herbert H. Clark and Edward F. Schaefer. 1989. Contributing to discourse. *Cognitive Science*, 13:259–294. Also appears as Chapter 5 in (Clark, 1992).

Herbert H. Clark. 1992. *Arenas of Language Use.* University of Chicago Press.

Herbert H. Clark. 1994. Managing problems in speaking. *Speech Communication*, 15:243 – 250.

Nils Dahlbäck and Arne Jönsson. 1998. A coding manual for the linköping dialogue model. unpublished manuscript.

Pierre Dillenbourg, David Traum, and Daniel Schneider. 1996. Grounding in multi-modal task-oriented collaboration. In *Proceedings of the European Conference on AI in Education.*

Discourse Resource Initiative. 1997. Standards for dialogue coding in natural language processing. Report no. 167, Dagstuhl-Seminar.

A. I. Goldman. 1970. *A Theory of Human Action.* Princeton University Press, Princeton, NJ.

Barbara J. Grosz and Candace L. Sidner. 1986. Attention, intention, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Peter A. Heeman and James Allen. 1994. The TRAINS 93 dialogues. TRAINS Technical Note 94-2, Department of Computer Science, University of Rochester.

Peter A. Heeman and James F. Allen. 1995. The Trains spoken dialog corpus. CD-ROM, Linguistics Data Consortium, April.

Julia Hirschberg and Christine Nakatani. 1996. A prosodic analysis of discourse segments in direction-giving monologues. In *Proceedings of the 34th Annual Meeting of the ACL*, Santa Cruz. Association for Computational Linguistics.

Diane J. Litman, Shimei Pan, and Marilyn A. Walker. 1998. Evaluating response strategies in a web-based spoken dialogue agent. In *Proceedings COLING-ACL-98.*

Karen Lochbaum. 1994. *Using Collaborative Plans to Model the Intentional Structure of Discourse.* Ph.D. thesis, Harvard University. Available as Technical Report 25-94.

Christine H. Nakatani and David R. Traum. 1999. Coding discourse structure in dialogue (version 1.0). Technical Report UMIACS-TR-99-03, University of Maryland.

Christine H. Nakatani, Barbara Grosz, David Ahn, and Julia Hirschberg. 1995. Instructions for annotating discourse. Technical Report 21-95, Center for Research in Computing Technology, Harvard University, Cambridge, MA, September.

Rebecca Passonneau and Diane Litman. 1997. Discourse segmentation by human and automated means. *Computational Linguistics*, 23(1):103–140.

Martha E. Pollack. 1986. *Inferring Domain Plans in Question-Answering*. Ph.D. thesis, University of Pennsylvania.

S. Siegel and N. J. Castellan. 1988. *Nonparametric statistics for the Behavioral Sciences*. McGraw-Hill, 2nd edition.

David R. Traum. 1994. *A Computational Theory of Grounding in Natural Language Conversation*. Ph.D. thesis, Department of Computer Science, University of Rochester. Also available as TR 545, Department of Computer Science, University of Rochester.

David R. Traum. 1998. Notes on dialogue structure. Unpublished manuscript.