The Report of The Third Workshop of the Discourse Resource Initiative, Chiba University and Kazusa Academia Hall

Mark Core Masato Ishizaki Johanna Moore
Christine Nakatani Norbert Reithinger
David Traum Syun Tutiya

Contents

Acknowledgments

Masato Ishizaki and Syun Tutiya

The workshop organizers are glad to acknowledge the generous financial support from Nissan Science Foundation and Chiba University, which made the workshop possible.

The graduate and undergraduate students of Chiba University who helped us organize the tutorials and workshop are to be thanked. Special thanks are to go to Mika Enomoto, who patiently followed the wild and winding discussions in the workshop, preparing the minutes. Toshiyuki Kawashima, who maintained the computer and network environment in Kazusa, will be fondly recalled by email addicts. Junko Arao and Eri Ishizaki deserve hearty gratitude from all the participants though they were in the backstage, for they were responsible for most of the secretarial chores, from sending and receiving mail, to listing the participants, to accounting, to photocopying, and finally to babysitting.

The tutorial presenters should be applauded for their impressive talks, which convinced the Japanese audience of the importance of collaborative corpus building efforts.

The contributers to this report are to be thanked by the editors and workshop organizers for their efforts in reconstructing what was going on in Chiba in May, 1998. Norbert Reithinger deserves special praise both for his having inadvertently volunteered to edit this report and now keeping his word.

All the people involved in this workshop and report writing are to be remembered for their contribution to the advancement of discourse and dialogue research in worldwide cooperation.

Some Lines from the Final Editor

Norbert Reithinger

I'd like to thank Syun Tutiya, Masato Ishizaki, Johanna Moore, and David Traum, who compiled all the texts in this report and all the other contributors who provided them with input. My task was finally just to do some formatting and provide basic consistency of the different parts.

List of Contributors

Masahiro Araki	University of Kyoto
Ellen Gurman Bard	University of Edinburgh
Jean Carletta	University of Edinburgh
Jennifer Chu-Carroll	Bell Labs.
Mark Core	University of Rochester
Morena Danieli	CSELT
Yasuharu Den	Nara Advanced Institute of Science and Technology
Barbara Di Eugenio	University of Illinois at Chicago
Mika Enomoto	Chiba University
Peter Heeman	Oregon Graduate Institue of Science and Technology
Julia Hirschberg	AT&T Labs.
Akira Ichikawa	Chiba University
Masato Ishizaki	Japan Advanced Institute of Science and Technology
Hideki Kashioka	ATR Interpreting Telecommunications Research Labs.
Masahito Kawamori	NTT Basic Research Labs.
Hideaki Kikuchi	Waseda University
Koiti Hasida	ElectroTechnical Lab.
Yasuo Horiuchi	Chiba University
Toshihiko Itoh	Toyohashi Univeristy of Technology
Susanne J. Jekat	University of Hamburg
Dan Jurafsky	University of Colorado
Yasuhiro Katagiri	ATR Media Integration and Communications Research Labs.
Hanae Koiso	National Language Research Institute
Tomoko Kumagai	National Language Research Institute
Lori Levin	Carnegie Mellon University
Diane Litman	AT&T Labs.
Kikuo Maekawa	National Language Research Institute
Johanna Moore	University of Edinburgh
Christine H. Nakatani	Bell Labs.
Shu Nakazato	Meio University
David G. Novick	Eurisco
Owen Rambow	CoGenTex
Norbert Reithinger	DFKI
Teresa Sikorski	University of Rochester
Michael Strube	University of Pennsylvania
Masafumi Tamoto	NTT Basic Research Labs.
David Traum	University of Maryland
Syun Tutiya	Chiba University
Jennifer J. Venditti	Ohio State University
Yoichi Yamashita	Ritsumeikan University
Hiroyuki Yano	Communition Research Labs.
Takashi Yoshimura	Electrotechnical Lab.
Gregory Ward	Northwestern University

1 The Plan of the Workshop

Masato Ishizaki and Syun Tutiya

1.1 Background

The third international workshop for discourse research initiative was held at Chiba, Japan from May 18 to 22, 1998, as a sequel to the first workshop in Pennsylvania, USA, in March 1996 and the second in Schloß Dagstuhl, Germany, in February 1997. The first workshop revealed the shared interest in exchanging the information on the existing discourse annotation schemes for illocutionary acts, discourse structure and co-reference and, in addition, discussed such issues as segmentation of discourse for analysis and the availability of computer tools. The second workshop examined the homework by the committed participants of applying the results of the first workshop to a limited but significant set of dialogues, focusing on the issues like categories, units and tools, and finally proposed a tentative annotation scheme which was called DAMSL (Dialog Act Markup in Several Layers), as stated in the report from the second workshop [Carletta *et al.*, 1997a, Allen and Core, Draft 1997].

The first and second workshops were held on invitation basis. That was because the workshops were intended to produce concrete proposals and standards for discourse and dialogue research and data sharing. In this sense, the workshops were a sort of committee meetings with a definite view to preparing grounds for resource sharing in relevant fields, and in fact they contributed to the progress of research on discourse to a great extent in 1996 and 1997.

Planning the third workshop, the local organizers thought that it was time to have a more or less commonly accepted tripartite format for a conference, namely tutorials, contributed papers, and committee meeting, since the outcomes of the two preceding workshops would deserve further dissemination beyond the original group and criticism and appraisal from the researchers of the similar interests. The lack of personnel and time, the shaky funding situation, and the almost simultaneously scheduled other conferences with the same research targets forced us to drop a session for contributed papers, leaving us only with an option consisting of a tutorial and a workshop.

1.2 Tutorials

The tutorials were planned to review the basics for designing annotation schemes, current standardizing efforts in the world, reports from ongoing research projects, and the advantage of having discursively annotated corpora in laying out the foundation of the empirical research as well as providing the material for discussing the future direction of research. What follows is a list of the tutorials:

- Masahiro Araki, Yoichi Yamashita, Shu Nakazato and Yasuo Horiuchi: The Current Status on the Standardization in Japan
- Norbert Reithinger: The Current Status on the Standardization in Europe

- Dan Jurafsky: Switchboard Spoken Language Modeling Project
- Julia Hirschberg: Prosody and Discourse Structure
- Jean Carletta: Discourse Analysis Reconsidered
- Ellen Bard and Jean Carletta: Exploiting the Characteristics of a Designed Corpus
- Barbara Di Eugenio and Jennifer Chu-Carrol: Learning from Annotated Corpora for Discourse Processing
- Diane Litman: Evaluating of Dialogue Systems
- Koiti Hasida: Annotated Texts as Versatile, Intelligent Contents

The tutorial sessions were held on Chiba University's Nishichiba campus on May 18 and 19 and participated in by more than 100 interested researchers and students, mostly from within Japan but some from abroad e.g. China and Korea. All the talks were very well received and induced a vigorous discussion. The intended purpose of dissemination was achieved.

1.3 The "Real" Workshop

From May 20 to 22, the "working" workshop was held at Kazusa Academia Hall, a pretty secluded place with a nice collection of conference equipments in Kisarazu area, 40 km south east of Chiba City. The following people attended and contributed to the workshop:

Masahiro Araki	University of Kyoto
Ellen Gurman Bard	University of Edinburgh
Jean Carletta	University of Edinburgh
Jennifer Chu-Carroll	Bell Labs.
Mark Core	University of Rochester
Morena Danieli	CSELT
Yasuharu Den	Nara Advanced Institute of Science and Technology
Barbara Di Eugenio	University of Illinois at Chicago
Mika Enomoto	Chiba University
Peter Heeman	Oregon Graduate Institue of Science and Technology
Julia Hirschberg	AT&T Labs.
Akira Ichikawa	Chiba University
Masato Ishizaki	Japan Advanced Institute of Science and Technology
Hideki Kashioka	ATR Interpreting Telecommunications Research Labs.
Yasuhiro Katagiri	ATR Media Integration and Communications Res. Labs.
Masahito Kawamori	NTT Basic Research Labs.
Hideaki Kikuchi	Waseda University
Akira Kurematsu	University of Electro-Communications

Koiti Hasida	ElectroTechnical Lab.
Yasuo Horiuchi	Chiba University
Toshihiko Itoh	Toyohashi Univeristy of Technology
Susanne J. Jekat	University of Hamburg
Dan Jurafsky	University of Colorado
Hanae Koiso	National Language Research Institute
Tomoko Kumagai	National Language Research Institute
Lori Levin	CMU
Diane Litman	AT&T Labs.
Kikuo Maekawa	National Language Research Institute
Johanna Moore	University of Edinburgh
Christine H. Nakatani	Bell Labs.
Shu Nakazato	Meio University
David G. Novick	Eurisco
Owen Rambow	CoGenTex
Norbert Reithinger	DFKI
Teresa Sikorski	University of Rochester
Michael Strube	University of Pennsylvania
Masafumi Tamoto	NTT Basic Research Labs.
David Traum	University of Maryland
Syun Tutiya	Chiba University
Jennifer J. Venditti	Ohio State University
Yoichi Yamashita	Ritsumeikan University
Hiroyuki Yano	Communication Research Lab.
Takashi Yoshimura	ElectroTechnical Lab.
Gregory Ward	Northwestern University

We called the efforts in the past 2 years `standardisation' or `standardization,' but that was a misnomer. In the typical standardising efforts, as seen in audio-visual and telecommunication technologies, companies try to expand the market for their products by making their products or interfaces as the standards for the purpose of making more and more profits. The objective of our efforts through the past workshops were to promote interactions among different groups working in discourse research and thereby provide firmer foundations for corpus-based discourse research, by saving researchers such wasteful duplicate efforts for the creation of resources, namely dialogue corpora of high quality. The series of workshops, we hope, will lead to the increase in number and quantity of the resources to be shared. The people involved in the workshops will be proud to call their efforts "data sharing" rather than "standardization."

The issues we addressed in this workshop were

1. Forward and backward communicative functions and
2. Common grounding units.

The participants of the workshop were assigned to either of the two discussion groups working on the two different, but related issues.

The decision on the agenda for this third workshop had a historical reason. That was due to the discussion after the second workshop in Schloß Dagstuhl, where there was a significant remodeling of what was formerly called illocutionary acts, and which, after that, were renamed forward and backward looking communicative functions. Some time after the second workshop, we had the first version of DAMSL annotation scheme and its computer tool, owing to the joint efforts by James Allen, Mark Core, Johanna Moore, David Traum and Peter Heeman. The proposed scheme captured the logical union of the essential insights displayed by most of the existing schemes such as those for MapTask (University of Edinburgh), Verbmobil (DFKI), University of Southwester Louisiana, and TRAINS (University of Rochester). However,

1. the agreement rate was yet to improve,
2. how to build up a higher level structure from the constituent units in terms of DAMSL categories was not clear,
3. the multi-lingual issue, i.e., how to apply the DAMSL scheme to other languages than English, was to be addressed and
4. rhetorical relations between utterances within a turn should be elaborated in more details as a first step for modeling the content of the utterances.

Apparently we would not have enough time to address all of the issues, and thus decided to focus mainly on the issues 1 and 2. The issue 1 concerned the classification of communicative acts and intentions. The session in this workshop for this problem was prepared and chaired by Johanna Moore & Mark Core. The issue 2 concerned a variety of ways of looking at the "structure of a dialogue," which have been proposed and discussed. To give a coherent understanding of the problem, David Traum & Christine Nakatani came up with a method of tagging dialogues based on the theory of common grounding in advance of the workshop and people were asked to homework and prepare. They co-chaired the discussion session.

The discussions in the workshop and afterwards are reported later in this brochure by the chairs of the discussion groups.

At the last plenary session, future plan was discussed and the participants agreed that they would pursue the further possibility of bringing back the fruits from this workshop together in the next workshop for discourse resource sharing.

2 Forward and Backward Communicative Functions

Johanna Moore

2.1 Members of the Backward- and Forward-Looking Group

Masahiro Araki	University of Kyoto
Jean Carletta	University of Edinburgh
Mark Core	University of Rochester
Morena Danieli	CSELT
Barbara Di Eugenio	University of Illinois at Chicago
Mika Enomoto	Chiba University
Akira Ichikawa	Chiba University
Toshihiko Itoh	Toyohashi Univeristy of Technology
Susanne J. Jekat	University of Hamburg
Dan Jurafsky	University of Colorado
Hideki Kashioka	ATR Interpreting Telecommunications Research Labs.
Masahito Kawamori	NTT Basic Research Labs.
Hideaki Kikuchi	Waseda University
Akira Kurematsu	University of Electro-Communications
Lori Levin	CMU
Johanna Moore	University of Edinburgh
Shu Nakazato	Meio University
David G. Novick	Eurisco
Norbert Reithinger	DFKI
Teresa Sikorski	University of Rochester
Hiroyuki Yano	Communication Research Lab.

2.2 Introduction

The group felt that we should begin by taking a step back and reminding ourselves of the research questions that we hoped could be answered and the applications that would be informed if we had large corpora tagged with the backward- and forward-looking annotation scheme (henceforth BF) under development. The results of this discussion are summarized in Section 2.3.

In order to prepare for the workshop, the group performed a homework assignment that involved coding three task-oriented dialogues with the DAMSL coding scheme, an annotation scheme that grew out of the two prior DRI workshops held at the University of Pennsylvania in March 1996 and Schloß Dagstuhl in February 1997.[1] A complete description of the homework assignment and the DAMSL coding manual that was used to annotate the homework dialogues may be found at `http://www.cs.rochester.edu/research/trains/-annotation`.

[1] For pointers to these prior workshops see `http://www.georgetown.edu/-luperfoy/Discourse-Treebank/dri-home.html`, `http://www.dfki.de/dri`, and `http://www.cs.rochester.edu/research/trains/annotation/`.

Each participant submitted annotated dialogues electronically to Mark Core at the University of Rochester. Using a suite of tools developed at Rochester, he computed agreement (using κ) for all coding categories. In order to identify specific problems that our group should address, we reviewed the results of the homework assignment, focusing on those categories that were frequent in the data, and for which intercoder agreement was low.[2]

To further elucidate problematic issues, we analyzed a dialogue from the Verbmobil corpus as a group, discussing utterances where there was considerable disagreement and arriving at a consensus coding. These exercises led to a series of specific tasks that were then assigned to smaller working groups. Section 2.4 describes the issues that the working groups addressed, and the sections that follow summarize the results of each working group.

2.3 Research Questions and Applications

There was overwhelming consensus in the group that a large corpus (reliably) annotated with dialogue acts of the type used in the BF scheme could shed light on several research questions of importance to the dialogue community. These include:

- In very general terms, how do dialogue participants negotiate and come to agreement? More specifically, how do they signal (mis)understanding, (dis)agreement, clarification subdialogues, and so on, with (sequences of) dialogue acts?
- In spoken dialogue, what, if any, is the correspondence between prosodic features (e.g., pitch and duration) and dialogue acts?
- What are the lexical and/or syntactic patterns for realizing different dialogue acts?
- How are dialogue acts realized in different languages?

There are many applications for which a tagged corpus would be useful. We identified the following:

- Genre detection
- Summarization
- Machine Translation
- Training classifiers
- Building dialogue systems

[2]The results of the homework assignments may be found at `http://www.cs.rochester.edu/research/trains/annotation/results.html`.

2.4 Issues Identified by FB Group

The FB group identified several issues that should be addressed at this or a subsequent workshop. These included several "large" issues, many of which were not addressed at this meeting, and several more tractable issues, some of which became the subject of small working groups.

At the workshop, subgroups focused on the following issues, and more detail about progress on these issues is provided in subsequent sections:

- Intercoder reliability for the **Statement** category was very low, with $0.33 < \kappa < 0.55$. Group members identified two possible problems: (1) the test for whether an utterance is a statement was confusing, and (2) distinguishing between **Assert** and **Reassert** was also a source of unreliability.

- Agreement on the binary decision about whether something was an **Info-request** was also low, $0.58 < \kappa < 0.62$. After examining several problematic examples, the group decided that introducing a tag for **Check** questions would be helpful. Section 2.6 addresses this issue.

- The **Answer** dimension was also problematic. Coders were not sure whether this dimension only included responses to questions or whether other things could be tagged as **Answer**. In addition, some suggested that we should further refine the **Answer** category, while others suggested that it be merged with **Agreement**. Section 2.7 addresses this issue.

- We revisited the question of where to draw the line and say that the hearer "understands" an utterance. There was also a question about whether one could tag anything other than **Hold** in the **Agreement** dimension if they had tagged the utterance as **Signal non-understanding** along the **Understanding** dimension. Section 2.8 addresses this issue.

- Some problems with agreement in the **Influence-on-Listener** dimension arose because coders had difficultly deciding between **Open-Option** and **Action-Directive**. The decision tree for this dimension was revised. Section 2.9 addresses this issue.

- We revisited the question of low-level segmentation. Section 2.10 addresses this issue.

- In discussions of why the reliability of the DAMSL coding scheme was so poor, the question of whether there is something inherently problematic about multi-dimensional coding schemes arose. We discussed "flattening" the categories as was done for the SWBD-DAMSL scheme, and also whether we really need all the dimensions in DAMSL. Jean Carletta argued that the style of the coding manual/scheme was crucial and that the current DAMSL manual/scheme was unwieldy. There was considerable discussion about this point and we return to it later. Section 2.12 addresses this issue.

The following are issues identified by the group, but that we did not have time to address at Chiba. We hope that some of these could be addressed at subsequent meetings.

- How does the BF coding relate to the higher-level discourse structure coding being done by the other group? In particular, our "Understanding" dimension seems to overlap considerably with the CGU's of that group.
- How could or should we code the strength of beliefs or commitment of the dialogue participants? For example, "suggestions" can be viewed as a weak form of "command". Should we be trying to capture this? Do we want to indicate varying strengths of a single category?
- How could or should we deal with utterances that seem to be intentionally ambiguous?
- How should we code disfluencies?
- What, if any, is the forward-looking function of an acknowledgment?
- How should we deal with non-verbal information, such as gaze (where video is provided), hesitation, silence, and timing?

2.5 Statements

By Lori Levin, Dan Jurafsky and Norbert Reithinger

The decision tree for statements in the DAMSL manual caused confusion because the top-level node asked the question "Does the speaker make a claim about the world?" Several coders argued that if the speaker was making a claim about his/her beliefs or the beliefs of another person, then the answer to this should be "no", and thus they did not code utterances about beliefs as **Statement**s.

The group decided to remove the phrase *about the world* from the top-level node of the statement decision tree.

The revised decision tree is as follows:

Does speaker make a claim?
- **Yes**. Tag as **Statement**
  - Does speaker think claim already made?
    - **Yes**. Tag as **Reassert**
    - **No**. No tag.
- **No**. No tag.

We agreed that any further refinement along the statement aspect would be up to the discretion of individual projects.

2.6 Check Questions

By Dan Jurafsky

Check is a new dialog act, a subtype of **Information-Request**. A check requests the listener to confirm information that the listener has privileged knowledge about. Usually the speaker has some reason to believe the information, but isn't sure about it. In task-oriented dialog the information usually comes from the preceding dialog. For example Carletta et al. [1997b] suggest that typically in the Map Task either the interlocutor may have tried to convey the information explicitly, or the speaker may believe the interlocutor meant to it be inferred from what the interlocutor said, as in 1:

(1) Map Task

G. . . . you go up to the top left-hand corner of the stile, but you're only about a centimeter from the edge, so that's your line.

F. OK, **up to the top of the stile?**

But especially in non-task oriented dialogue checks may also be used to address subjective sentiments of the interlocutor, or the interlocuter's opinion:

(2) Nightline news interview [Heritage and Roth, 1995]

IR. You agree Senator that whether anybody likes it or not Central America is a shadow in all of this?

F. .hh Well of course e::h it's important to our interests and . . .

Checks are equivalent to the **Check** move of HCRC's Map Task scheme [Carletta *et al.*, 1997b], SWBD-DAMSL's **Reformulate** tag [Jurafsky *et al.*, 1997], and what [Labov and Fanshel, 1977, 100] have called **requests for confirmation**. They subsume what Labov and Fanshel have called **B-event** statements (see also Pomerantz, 1980), and checks overlap significantly with Verbmobil's **Request-Clarify** tag [Alexandersson *et al.*, 1997].

Checks in English may be realized as tag questions [Quirk *et al.*, 1985, 810-814], as declarative questions with rising intonation [Quirk *et al.*, 1985, 814], as declarative questions without rising intonation, and as `fragments' (subsentential words or phrases) with rising intonation. Checks seem to be most commonly realized by the declarative structures; we have no examples of checks realized as Yes/No questions with aux-inversion [Quirk *et al.*, 1985, 807-810] although presumably this is possible.

However they are realized, Checks tend to be responded to by the listeners as if they were either yes-no questions or proposals to be accepted or rejected, as in the following example from therapeutic discourse.

(3) [Labov and Fanshel, 1977]

Th. And it never occurred to her to prepare dinner.

R. No.

Th. She was home all afternoon.

R. No, she doesn't know how.

Th. But she does go to the store ...

R. Yes.

The therapist repeats or extends information that the patient R has given. Note that the therapist's statements are all in declarative form; they have the superficial appearance (the locutionary force) of statements. But their illocutionary force here is that of a question, as we can see from the `Yes' and `No' responses.

Labov and Fanshel propose the Rule of Confirmation to account for the way these utterances function as questions. The Rule of Confirmation is based on the following classification of statements according to how knowledge is shared among the two participants (page 100):

```
A-events:   Known to A, but not to B.
B-events:   Known to B, but not to A.
AB-events:  Known to both A and B.
O-events:   Known to everyone present.
D-events:   Known to be disputable.
```

A-events may concern A's past history, beliefs, emotional state, and so on. The classification is based on agreements between the participants; if the classification of an event is disagreed upon, it falls into the class of D-events.

Labov and Fanshel state their Rule of Confirmation as follows:

> **Rule of Confirmation:**
> If A makes a statements about B-events, then it is heard
> as a request for confirmation.

2.6.1 Realizations of Check

Here are examples of each of the possible realizations of Checks in English, and a few examples in other languages.

1. Tag questions formed by an auxiliary and a subject (in that order) placed after a statement [Quirk *et al.*, 1985, 810-814]

   (4) Switchboard [Godfrey *et al.*, 1992]

   B. They just, they just announced that, didn't they?

   A. Yeah.

2. Tag questions formed with an `invariant tag' (usually `right')[Quirk *et al.*, 1985, 814]

   (5) Trains [Allen and Core, Draft 1997]

   U. and it's gonna take us also an hour to load boxcars right
   S. right

3. As declarative questions with rising intonation [Quirk *et al.*, 1985, 814]

   (6) Switchboard [Godfrey *et al.*, 1992]

   A. and we have a powerful computer down at work.
   B. Oh (laughter)
   B. so, you don't need a personal one (laughter)?
   A. No

4. As declarative questions without rising intonation.

   (7) Labov and Fanshel (1977)

   Th. But she does go to the store . . .
   R. Yes.

   They tested this rule in a series of interviews about life about New York City. If the subject reported a burglary, the interviewer inserted the following statement with declarative intonation:

   (8) Labov and Fanshel (1977)

   a. And you never called the police.

   All subjects responded as if it this had been a yes-no-question (of the form `And is it true that you never called the police?")

   (9) Map Task [Carletta *et al.*, 1997b]

   G. Right, em, go to your right towards the carpenter's house.
   F. Alright well I'll need to go below, I've got a blacksmith marked.
   G. Right, well you do that.
   F. Do you want it to go below the carpenter? [*]
   G. No, I want you to go up the left hand side of it towards green bay and make it a slightly diagonal line, toward, em sloping to the right.
   F. **So you want me to go above the carpenter**? [**]
   G. Uh-huh.
   F. Right.

5. As fragment questions (sub-sentential units; words, noun-phrases, clauses) [Weber, 1993]

(10) Map Task [Carletta *et al.*, 1997b]

G. Ehm, curve round slightly to your right.
F. **To my right?**
G. Yes.
F. **As I look at it?**

(11) Switchboard

B. And what do you think you'll do with that?
A. **With those degrees?**
B. Uh-huh.

6. German example

(12) Verbmobil [Alexandersson *et al.*, 1997]

Montag, Dienstag, sagen Sie, neunzehnter, zwanzigster Juni?
Monday, Tuesday, say You, nineteenth, twentieth June?

7. Japanese example

(13) Verbmobil [Alexandersson *et al.*, 1997]

ku	gatsi	no	nanoka	desu	ka.
Card	N	Part	Date	V	Part
9	month	(Genitive)	7th	be	(Question)

"Is this September the 7th?"

2.6.2 Backward Function of a Check

Checks have the backward-looking function **Signal-non-understanding**. This is because currently we only have 2 classes of backward-looking functions with respect to understanding: **Signal-understanding** and **Signal-non-understanding**. Thus, it was deemed unnecessary to add a third `in-between' category for confirmation-type dialogue acts like **Check**.

2.6.3 Responses to Checks

Heritage and Roth (1995) state that a check "makes a recipient's confirmation or denial relevant in the next turn". This suggests that responses to checks should be coded along the **Accept** dimension (**Accept**, **Partial-Accept**, **Partial-Reject**, Reject). Allen and Core (1997) suggest that the response to a check also be coded as **Answer**.

In normative cases, then, we suggest that checks be coded as both **Answer** and also along the Accept/Reject dimension. Individual projects may choose to have coders only mark one of these responses. for example SWBD-DAMSL coded responses to **Reformulations** only as **Accept** or **Reject**; all responses to **Reformulations** were considered to be implicitly **Answers**, and so did not need to be explicitly coded as such.

Here is a repeat of example 6 above showing the response as an **Accept** (recall that "no" is often used to accept a negative question or proposition:)

(from 6) Switchboard

CHECK B:. so, you don't need a personal one (laughter)?
ACCEPT A:. **No**

(from 10) Map Task

CHECK F:. To my right?
ACCEPT G:. **Yes.**

(14) Switchboard

CHECK A:. Joe Pros?
REJECT B:. Uh, they're not Joe Pros

In general the distribution of answers types to checks looks like a mix of answers to yes-no questions and accept/rejects of proposals. Here are examples from the SWBD-DAMSL database of the responses to 800 **Reformulations**, a subtype of **Check**:

%	Count	Response to Checks
35%	281	yeah
10%	83	right
6%	49	uh-huh
6%	44	and...
3%	21	yes
3%	20	no
2%	18	that's right...
1%	9	okay
1%	8	exactly
1%	7	oh yeah
1%	5	well yeah
32%	255	[other]

Sometimes an **Accept** response is realized as shared laughter. In the following segment, B.3 functions as a check and A.4 as its response; A then goes on in A.5.

(15) Switchboard

A.1. . . . and then after one year it started heating only on one side.
B.2. Huh,
B.3. so either you or your husband can be warm but not both (laughter).
A.4. (laughter)
A.5. Also, I took an iron back after having it only one year.

2.7 Answer vs. Agreement

By Masahiro Araki, Mark Core, and Mika Enomoto

At the Dagstuhl meeting, the **Answer** dimension of the backward-looking communicative function was considered problematic except for the clear case in which the utterance was only a response to a question. For example, one problematic case is utterance 3 of Example 16, which is an answer to a question involving an action. In DAMSL utterances are allowed to simultaneously accept, answer, and commit so an annotator must consider what else besides an answer utt3 is.

(16) utt1. s: oh you need a boxcar to carry oranges
utt2. do you want to pick one up at Dansville?
utt3. u: yes

Unifying the **Answer** and **Agreement** dimensions would eliminate the confusion about whether an answer is also an acceptance. There would still have to be a category for examples such as utterance B:

(A) u: how long is that?

(B) s: two hours

which simply answers a wh question, but acceptances and positive question answers would be unified into one category and rejects and negative question answers would be unified into another.

This approach is motivated by the Japanese Discourse TAGging scheme (JDTAG), which takes a different view of backward-looking functions. In JDTAG, in principle, each utterance is coded using a single utterance unit tag. The exception is utterances that have a function of "respond with initiating", i.e., utterances that have one backward function and one forward function. The typical cases of "respond with initiating" are holds and questions, which begin an embedded clarification sub-dialogue. In the JDTAG scheme, an utterance such as utt3 above would simply be coded as a positive response without the need to consider whether it also accepts or commits.

Example 17 is a problematic example for any scheme since it is tempting to label utterance B as "answering" utterance A; although since utterance A explicitly states that the speaker does not want to know the current time, it should not be labeled as an answer. Some workshop participants believe it is easier to handle in JDTAG because once the annotator decides that utterance B has a forward function then the possibility of it being an answer can be eliminated. This would be further evidence for the advantages of a single or two dimensional scheme.

(17) A. I should be at a meeting. Luckily, I don't know what time it is.
B. It's 3 o'clock.

Further compressing the dimensions of DAMSL so that it more resembles JDTAG brings up the problem of making sure the new scheme can capture all the phenomena that were marked in DAMSL by combinations of tags. The following paragraphs report on the results of the JDTAG working group on trying to map the answer/agreement dimensions of the DAMSL scheme to the one-dimensional JDTAG scheme.

There are problematic cases where it is difficult to decide whether the answer is positive or negative. For example, whether the response in 18 is tagged as positive or negative depends on whether we consider something that is a 10 minute walk away to be near or not.

(18) A. Is the hall near the station?

B. It takes about ten minutes walk.

Another problematic case occurs in the confusion of the surface expression and the content of the utterance. One usage of the Japanese *hai* is affirmation of the content of the previous utterance. Some coders who consider such a *hai* as an affirmation may code B in Example 19 as positive/agree. But others may code B as negative/reject, considering its content.

(19) A. hoka ni yotei ha arimasen ka
Aren't there any other plans?
B. hai arimasen
No, there aren't.

As a result, we can say that the obligatory multidimensional tagging scheme yields problematic cases in some general dimensions. A one (or at most two) dimensional tagging scheme may be one solution to this problem. However, the multidimensional tagging scheme has the advantage of capturing utterances that apply to more than one tag of the scheme. If we wish to continue using a multidimensional tagging scheme, then each dimension should be orthogonal to the others. That is, coders must be able to decide on the classification of each dimension without considering any another dimension. If dimensions are not orthogonal, the manual should list the restrictions on selecting classifications.

2.8 Understanding

By Mark Core and David Novick

2.8.1 Current Categories

Signaling understanding is a topic that got considerable discussion at Dagstuhl and that requires considerably more work and collaboration with the group looking at Discourse Structure. Recall that the understanding dimension of the BF tagging scheme deals with what the current utterance says about the participants ability to recover the explicit content of the antecedent.

The current categories are:

1. **Signal non-understanding**: utterance indicates that participant could not recover the explicit meaning of the antecedent. Examples:

   (20) Huh?
   What?
   What did you say?
   I didn't understand you.

2. **Signal understanding**: the utterance indicates that the participant was able to recover the explicit meaning of the antecedent. There are several types of confirmation behavior that fall into this category. We break this category down into the following subcategories:

   - backchanneling
   - acknowledgments
   - repetition/rephrase
   - completion

   In general, if an analyst explicitly tags an utterance at the agreement level, it implies a **signal understanding** tag at the understanding level.

   Completions. These are when one speaker starts to perform an illocutionary function, and it is completed by another speaker. A prototypical examples is:

   (21) A: A train will arrive at ...
   B: 4pm

3. **Re-realize**: This category covers cases where the utterance corrects misspeaking in the antecedent.

   (22) A: Let's take engine E2.
   B: You mean E1.
   A: E1 to Dansville.

2.8.2 Open Problems

Completions At Dagstuhl, we decided that we needed to isolate examples of these collaborative completions (such as the following) from our corpora to see what's really involved with tagging them.

(23) A: The train arrives at # .... #
B: # 4pm on #
A: Friday

Everyone should look for these kinds of examples in their corpora, and present examples to the group to see what's involved with coding them. For some domains (e.g., translation domains such as Verbmobil), the problem might not arise.

What does it mean to recover the previous utterance(s)? At both Dagstuhl and Chiba, there were lengthy discussions about "where to draw the line" and say that the hearer has "recovered" the explicit semantic content of the antecedent. Consider the following responses to A:

(24)
A	Take some oranges to Dansville.
B_1	Huh?
B_2	I don't understand.
B_3	To Dansville?
B_4	Dansville New York?
B_5	Can I use a train to do that?

Proposals for criterion about where to draw the line were:

1. Response indicates that hearer can fully identify the speaker's intention in uttering the antecedent. The group did not like this criteria because we felt it required too much subjective reasoning about intention and the recognition of intention.

2. The acceptance test: Could an indication of acceptance have been included in the response? For example:

   (25) A: Take some oranges to Dansville.
   B: OK I will. Can I use a train to do that?

   sounds fine, whereas

   (26) A: Take some oranges to Dansville.
   B: OK I will. Dansville New York?

   sounds odd.

   Using this criterion, B_1–B_4 would all be marked as **signal non-understanding**, and B_5 would be marked as **signal understanding**.

3. Has the hearer correctly identified the senses of the lexical items and the roles that constituents play in the semantic structure? Under this proposal, recovering the semantic content implies that no lexical or syntactic ambiguities remain, but does *not* imply that the hearer was able to resolve referents.

   Using this criterion, B_1–B_3 would all be marked as **signal non-understanding**. B_4 and B_5 do not indicate any evidence of non-understanding. In B_4, the hearer has heard the correct words, but can't perform reference resolution.

This was one of the main discussion points at the meeting. We originally opted for criterion 3, but then in a joint meeting with the forward-looking group reconsidered and opted for criterion 2. One concern is that it be easy to write clear instructions that enable coders, who are not necessarily trained in linguistics, to reliably code the data. In the meeting, we found criterion 3

difficult to explain without resorting to terms like "reference resolution" and "co-specification". Criterion 2 seems to have a simple test, but we will have to see how reliable the coding is on future homeworks.

2.9 Directives

By Barbara Di Eugenio and Morena Danieli

At the meeting in Dagstuhl, a distinction between two different kinds of Directives, **Open-Option** and **Action-Directive**, was introduced. In the Dagstuhl report, **Open-Option** was defined as applying to those cases in which the speaker S presents an option to the hearer H that S doesn't necessarily endorse — in the following excerpt, `[a]` is such an option, as made clear by `[b]`:

(27) A: We could buy my red sofa for $300,
B: however yours for $250 is probably better.

However, in the DAMSL draft, the notion of endorsement is not mentioned; rather, **Open-Option** is defined as "suggesting a course of action but without putting any obligation on H". This vague definition for **Open-Option** made it very difficult for coders to distinguish **Open-Option** from **Action-Directive** (in the following, **OO** and **AD** respectively). In fact, it turned out that coders would often resort to surface form to draw this distinction: a more directive form such as `[a]` in excerpt 27 would be tagged as an **AD**, a declarative form such as `[b]` would be tagged as an **OO**. Among other difficulties mentioned in the discussion were:

1. *Politeness and indirectness*. Languages differ in social conventions, and in particular in levels of politeness and indirectness. How do politeness and indirectness affect the distinction between **OO** and **AD**? For example, if in the US a professor and a student are scheduling a meeting together, both of them will probably utter many **AD's**; however, if the same conversation were taking place between a Japanese professor and student, in particular a female student, the student would use only **OO's**, not **AD's**.

2. *Lack of necessary information.* Sometimes even if the directive is given in an imperative form, as in (28) below, there is crucial information missing, so that S cannot really consider H under obligation to perform the corresponding action. For example, in the furniture purchasing domain, the price of each item must be known before a decision can be taken: thus, after 28 is uttered, if H doesn't know the price of S's sofa, H cannot be assumed to be under obligation to purchase that sofa.

   (28) Let's buy my red sofa.

To address these problems, we propose the decision tree in Figure 1. The following comments refer to the labels of the nodes in the tree. First, we discuss the tests in the internal nodes.

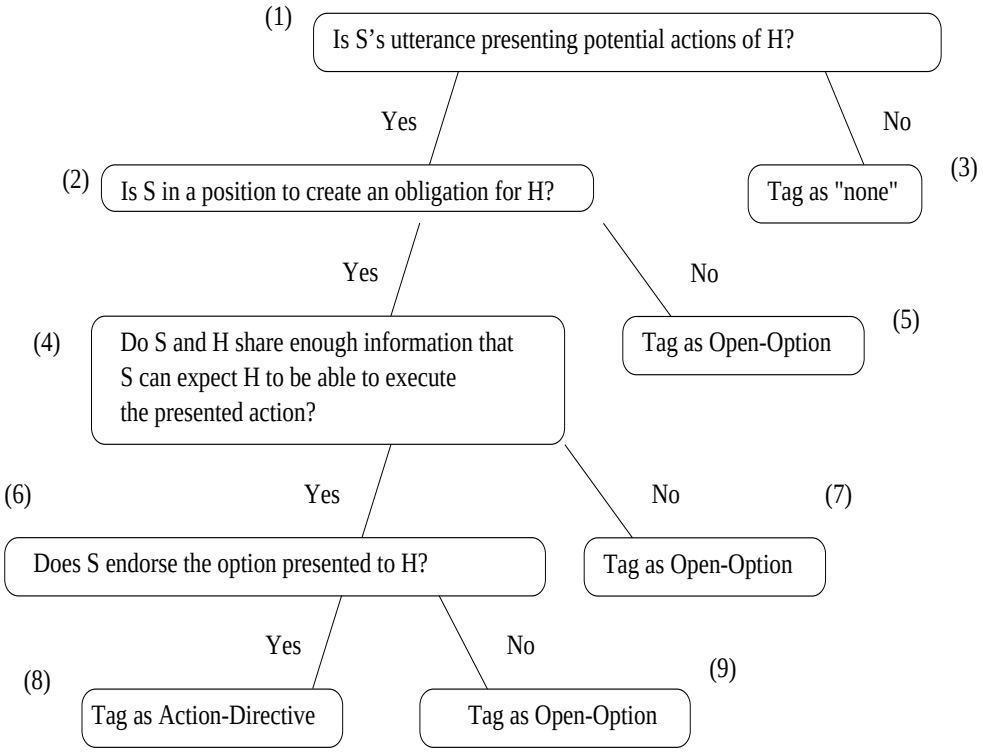


Figure 1: Decision tree for Open Option and Action Directives

(1) We use the term "presenting" as a neutral reference to the content of S's utterance. Notice that "potential actions" need to be defined for coders to be able to recognize them. We feel that such definition belongs to individual projects, not to the high-level coding document that DRI will put forward; however, this document should definitely mention the issue of defining what a "potential action" amounts to.

(2) This test covers those situations in which one speaker cannot create obligations for the other: e.g., a student in Japan, a subordinate in a company or in the army, a system in a **master-slave** configuration, etc.

(4) This test covers situations in which the speakers are collaborating on a problem, and part of the information is private. For example, assume that two speakers are trying to schedule a meeting. If S proposes a certain time, even if in a strongly directive form such as **Let's meet on Tuesday**, without knowing whether H is actually free on Tuesday, the utterance cannot be considered as an **AD**. Again, the test is vague because each project will have to define what *sharing enough information* means. Partly, the definition will depend on social conventions. For example, if the boss is trying to schedule a meeting with one of his employees, an utterance such as *Let's meet on Tuesday* on the part of the boss will count as an **AD** even if the boss doesn't know the employee's schedule: the hierarchical structure is such that the issue of the employee's own schedule does not

arise.

(6) This test captures those cases in which S makes it clear that the option is not necessarily endorsed, such as `[a]` in excerpt 27 above.

Regarding the leaves of the decision tree, we mention here some more specific labels that have been proposed at the meeting. It is debatable whether these more specific labels should be adopted by DRI, or chosen by individual projects only.

(5) There was a strong consensus at the meeting that in this case the label should be **Suggest**, rather than the more vague *Open-Option*. In fact, it was pointed out that this is the way *Suggest* has been used in many coding schemes, e.g. in Verbmobil.

(7) It was suggested that the label in this case could be **Weak Suggest** or **Negative Suggest** (alternatively, **Weak** or **Negative Open-Option**).

2.10 Segmentation of Okay-like Utterances in Speech

By Susanne J. Jekat and Teresa Sikorski

In this section we take a closer look at segmentation procedures applied to discourse data which is prepared for annotation. Principally, two different procedures are applied:

a) the whole turn of one speaker is annotated,

b) some turns are pre-segmented on the basis of prosodical, functional or structural criteria.

We will argue that only procedure a) is an adequate preparation for further analysis of functional characteristics of the data and that segmentation should be effected either simultaneously or after the analysis by annotation. The advantage of procedure a) is that the boundaries of a turn as a physically given unit are in most cases easy to recognize.

The problem of units that are "too large" seldom arises in the domain of task oriented dialogues which are considered here.

As for procedure b), there are several arguments against it:

1. Prosodic pre-segmentation is done while reading the transcript. This is no pure prosodic segmentation because the structure of the written texts comes into play.

2. At present, there is no general set of functional units applicable to a larger database and segmentation by functional criteria is influenced by individual or project specific definitions of functional units (Communicative functions in DAMSL, cf. [Allen and Core, Draft 1997]; Dialogue acts in Verbmobil, cf. [Alexandersson *et al.*, 1997]).

3. Structural criteria are derived from written language (e.g. a complete sentence consists of ....) and either do not fit into spoken texts or remain in an intuitive status while there is still no systematic definition of structural units like 'phrase' or 'utterance'.

If procedure b) is applied, pre-segmentation results in inconsistent data. The following examples demonstrating this are taken from the homework-dialogues for the third meeting of the Discourse Research Initiative in Chiba, May 1998 (Maptask, cf. [Carletta *et al.*, 1997b]; Verbmobil, cf. [Jekat *et al.*, 1997]) where 'okay-like-utterances' (OLUs in the following) are either separated from the rest of the turn and tagged as separate units or not.

Maptask

- Segment 1 - 2: OKAY.
  DO YOU HAVE A START X?
- Segment 8: OKAY, MY IRON BRIDGE IS RIGHT ABOVE THAT.
- Segment 49: SEE OKAY.

Verbmobil

- Segment 15 - 16: WELL
  WHEN ARE YOU GETTING BACK
- Segment 27 - 28: SOUNDS GOOD
  ON CAMPUS OR OFF
- Segment 32: SOUNDS GREAT EXCEPT THEY HAVE BEEN OUT OF BUSINESS FOR A WHILE

Different functions of OLUs are listed in figure 2, the list is not a complete one but should cover functions in task-oriented data like the above mentioned Maptask or Verbmobil.

The forward-looking function of OLUs in initial position is turn-keeping in different degrees ranging from non polite turn-grabbing to what we call here communicative function, which signals the speaker that the listener is still listening. OLUs within the turn seem to take their position between forward and backward-looking, as a cue to both repair (BW) and new start (FW) or topic shift (BW,FW).

The backward-looking function of OLUs in our corpora range from acoustic understanding that does not trigger any additional reaction in the hearer to several degrees of consequences as there are: questions concerning the signal (2b), verbal acceptance of a suggestion (2c), verbal acceptance plus grounding of the further communication on the acceptance (2d) and, verbal acceptance plus fulfilling the accepted action (2e).[3]

Tagging conventions of the DRI include the constraint that only the actual status of the discourse is to be tagged, that is taggers should not take into

[3]In Japanese, discourse particles like *hai* have an important function because they disambiguate neutral statements to be rejects or accepts.

Realisation of okay-like-utterances	Examples	Specific Function	General Function
in initial position	**English** yes well okay sounds good	1) turn-keeping a) turn-grabbing b) turn-taking c) communicative function d) repair	FW FW, BW
between turns	**French** oui d'accord	e) turn-yielding 2) Understanding	FW, BW BW
at the end of the turn	**German** okay ja gut **Japanese** hai **multilingual** aha hm	a) acoustic understanding b) acoustic understanding + consequence c) acoustic understanding + accept d) acoustic understanding + accept + consequence e) acoustic understanding + accept + action	

Figure 2: List of Phenomena

consideration the rest of the dialogue (no forward-looking without limits) but as shown in Table 1, there are many different functions of OLUs which very often can not be isolated on the basis of only the OLU item. We therefore claim, that OLUs should not be separated from the rest of the turn before further analysis (annotation) is done. Two reasons, at least, exist for this claim:

- Presegmentation is not based on consistent constraints,
- forward-looking is not allowed for taggers, but the rest of the actual turn can be taken into consideration if it is not presegmented from the OLU.

We expect better agreement on tagging of OLU-functions if the above mentioned procedure a) is applied for annotation.

2.11 Summary of the Discussion of the Japanese DTAG Subgroup

By Masahito Kawamori

The discussion of our subgroup centered around two main subjects: the sources of disagreement among coders of JDTAG and the possible connections between DAMSL coding scheme and JDTAG coding scheme.

2.11.1 Sources of Disagreement

The JDTAG coding scheme is based on the assumption, derived from observations from actual dialogues, that the main body of dialogue is made up sequences of three types of dialogue acts: initiate, respond, and follow-up (see Table 1). Such a sequence of acts is called an **IRF** sequence.

On the assumption that this IRF format is basically correct, the subgroup concentrated on the factors that seem to be most responsible for the disagreement among the coders. One factor that was pointed out was *aizuchi*, or **back-channels**, and many participants felt the need for some systematic way to tackle this phenomenon, like introducing a new category. Another factor also pointed out was systematically ambiguous expressions with polarized interpretations.

`Aizuchi' or Back Channels Our group agreed that one major source of disagreement among coders and of low α values is *aizuchi*. Although *aizuchi* func-

Type	Acts
Conventional	OPEN,CLOSE
Initiate	REQUEST, SUGGEST, PERSUADE, PROPOSE, CONFIRM, YES-NO Q, WH-Q, PROMISE, DEMAND, INFORM, OTHER ASSERTION, OTHERS
Respond	POSITIVE, NEGATIVE, ANSWER, HOLD, OTHER,
Follow-up	UNDERSTANDING

Table 1: Classification of acts in JDTAG

tions somewhat similarly to back-channels in English, many Japanese participants felt the need for introducing a dialogue act of just *aizuchi*. There are several reasons why *aizuchi* is responsible for low α values and why it is so difficult to handle.

First of all, it is generally acknowledged, as was emphasized by Professor Ichikawa, that *aizuchi* is multi-functional, or plain ambiguous. A good example is *hai*. *Hai* is one of the most frequently used words in spoken Japanese as well as one of the most complex and difficult to analyze. If used in response to a question, *hai* means a simple "yes", while if it is in reply to a request, it means an accepting "OK". When used by itself, at the beginning of a sentence, it usually corresponds to English discourse markers "now" or "well". In addition to all these functions, it also has its most common use as an expression of acknowledgment, as does "uh-huh" in English. It can be suspected that a Japanese speaker sometimes use *hai* intentionaly ambiguously; if this is the case it would be extremely difficult, if not impossible, to distinguish

It is to be noted also that *aizuchi* may also have a forward looking, as well as backward looking, function. For example, *nee*, which is usually understood as *aizuchi* by most Japanese, when interpreted as backward looking, would mean `isn't that so?', belonging to **Respond** or **Follow-up**; when it is interpreted as forward looking it means "don't you think so?" and should be categorized as **Initiate**.

The above discussion shows why *aizuchi* expressions are one of the major sources of disagreement among coders. One way to avoid this problem is to set up a category, as was suggested by the subgroup members, devoted to *aizuchi*. Dan Jurafsky also suggested that a **disjunctive category** can be introduced into the DRI general coding scheme, so that ambiguous utterances like *aizuchi* can be handled. We believe this is a very promising move.

Systematically Ambiguous Expressions As mentioned above, Japanese speakers can sometimes be very ambiguous, because of politeness or some other social constraints. Sometimes they use intentionally ambiguous expressions so that their true intentions are not, at least on the surface, known. An extreme case can be found in the way a rejection is conveyed.

Yoroshii-desu, which literally means `it is good', is often used as a polite form of rejection. The same can be said about the expression *Kekkoo-desu*, which is `That is OK'. The disturbing fact for the researcher is that since these expressions are used so that the fact the speaker is rejecting an offer is to be disguised, these expressions used as rejection and these expressions used literally usually have little prosodic differences. The decision as to which of the two alternative interpretations, NEGATIVE or POSITIVE, is to be taken has to be made entirely based on the context.

Such systematically ambiguous expressions may also suggest the need for the possibility of introducing a disjunctive category.

Units of tagging In the Chiba University Map Task project, any pause longer than 400 milliseconds is counted as delimiting a tagging unit. Similar meth-

ods based on properties of fundamental frequency and energy are adopted by many Japanese discourse resource projects. Such measures are not, as was noted by some participants, the results of extensive theoretical investigations but rather to be taken as expedient rules of thumb, however good heuristics they may be. This fact shows, the members agreed, how difficult it is to demarcate tagging units in spoken Japanese dialogue.

Some Japanese participants proposed that efforts should be made to start a serious attempt at reaching a general agreement as to what constitutes the most plausible measure of tagging units, not only pragmatically but also theoretically, and whether the conventional rules of thumb can be vindicated by such investigation.

2.11.2 Mapping Over the Coding Schemes

The DAMSL coding scheme has a multi-layer structure with roughly fifteen different dimensions of dialogue acts. JDTAG, on the other hand, is a single dimensional scheme with the above mentioned four major types of dialogue acts. One natural question to be asked is whether all the dimensions in DAMSL necessary, and another one is whether the dialogue acts in JDTAG sufficient. We mentioned above that the possible need for more categories by pointing out the cases of *aizuchi* and systematically ambiguous utterances.

Correlation among DAMSL dimensions Many of us thought that DAMSL coding scheme may be unnecessarily redundant; there may be some inherent implicational orderings among the dimensions or their components. It would be interesting to see if these possible implicational orderings can be attested empirically, using statistical methods, for example. Of course, a more rationalistic approach is also fruitful and exciting; for example, one can think of devising a calculus of implications over DAMSL dimensions, based on the decision trees given as definitions of these dialogue acts. The members of the subgroup believe that such efforts would clarify if DAMSL coding scheme is somewhat superfluous or too restrictive.

Comparison of DAMSL and JDTAG As mentioned above, the JDTAG coding scheme is based on uni-dimensional classification, and the direct comparison between the two schemes would be surely a difficult task. But we thought that work on the mappings between JDTAG and DAMSL should be undertaken, and in fact something similar to what we envisaged was already demonstrated at the workshop in the presentation by Masahiro Araki in his discussion of Answer, Agreement, and Understanding.

Such work, along with tasks like using JDTAG to annotate 'homeworks' for the workshop, would surely enlighten the usability of the coding scheme and clarify the similarities and differences between the two schemes, contributing much to the general utility of the DRI coding project as a whole.

2.12 Discussion of Coding Scheme/Manual Style

There was considerable discussion about whether the problems with low intercoder reliability were due to the multidimensional nature of the scheme, the unstated (and possibly unintended) interactions between dimensions, the sheer complexity of the scheme, or the quality of the coding instructions. Experience with SWBD-DAMSL has shown that higher intercoder agreement can be obtained by "flattening" the scheme into a uni-dimensional scheme and removing categories that correspond to combinations of tags that can never co-occur.

Jean Carletta also argued vehemently that the coding scheme should ideally be reduced to decision tree(s) that could fit on 1-2 pages.

The JDTAG group's experience also seems to indicate that a simpler, unidimensional scheme is both useful and can be coded reliably.

A vocal minority in the group argued that the multi-dimensional scheme was both easy to code and allowed the flexibility necessary to test the hypotheses of interest to their projects. Some members are of the opinion that this is an issue that depends from the purpose of the project, the type of coders used, etc. For example, if the project requires coding a lot of data using many coders that are inexperienced in language research, the uni-dimensional scheme may lead to better reliability and more efficient coding. For projects that require less coding using expert coders, the hierarchical scheme may be better. This is clearly a discussion that must continue.

In Chiba, the subgroup proposed developing a more general coding scheme (possibly uni-dimensional) that could serve as the "top level" for a range of more specific coding schemes that individual projects may want to devise. Specific coding schemes (be they multi- or uni-dimensional) should then be related to the general one. In this way, categories that are in the general scheme, and which are shared by many projects are separated from project specific requirements. Relations between general and specific schemes may then serve as basis for access to and evaluation of the project specific schemes.

2.13 Issues for Future Meetings

Remaining unanswered questions:

1. **What is being responded to?** We need a way to indicate what portion of prior discourse the current utterance responds to. This requires that we define the allowable scope of a response. At Dagstuhl, the backward looking group agreed that this issue depends on both the higher-level discourse structure and on decisions about the minimal unit of analysis (and thus hinges on decisions about segmentation).

   The group is looking at locality (typically within the previous utterance or turn), but we need further work to develop a precise definition of the allowable scope of a response.

2. **What constitutes the response?** A related question, which also initially arose at Dagstuhl, came from trying to apply DAMSL to other corpora.

At Dagstuhl, we found that in genres such as social conversation (e.g., phone conversations from the Switchboard corpus), it can be very difficult to determine where the "response" ends and where a new or sub topic starts. Again, the group felt that this was a very important issue, but that since it is really a higher level discourse structure issue, collaboration with that group would be necessary to make further progress on this issue. At Chiba, we focused our efforts on clarifying/refining the forward and backward tags, with the understanding that we expect further work on higher level discourse structure to be continued and eventually integrated with our work.

3. **Relation to higher level discourse structures?** At the Chiba meeting, there were some discussions about the relationship between the BF level of analysis and the analysis of CGU's. There seemed to be an question about whether the dialogue acts (BF functions) could be viewed as basic units that higher level discourse structures (HLDS) are composed of, or whether CGU's were the basic building blocks of HLDS. We clearly need further discussion to clarify the relationship between BF and CGU coding and their relationship to HLDS.

4. **How to integrate coreference coding with forward and backward communicative function?**. At the Chiba meeting there was no subgroup working on coreference. However, this is still an open question that must be discussed in the future.

5. Other tasks include

   - Specifying a set of informational relations (or at least some very general categories).
   - Coding of floor and topic control issues.
   - Coding of significant non-linguistic signals, e.g., refusing to respond, silence as acceptance.

3 Discourse Structure Coding

David Traum

3.1 Introduction

The general theme of the third DRI meeting was the analysis of task-oriented dialogue. Thus the discourse structure subgroup was concerned exclusively with the discourse structure of *dialogue*. This, in effect, called for a completely new group within the DRI framework, since previous DRI efforts concentrated only on discourse structure of monologue text, or individual dialogue acts. The discourse structure 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). Our first task was to circumscribe a set of phenomena that could be studied in detail leading up to (via participation in coding tasks) and during the meeting (in intensive working discussion sessions). The coding schemes, coding exercises and results are described in the next section. The activities at the meeting itself are then described in section 3.3.

3.2 Pre-meeting Activities

Finding a good starting point for a consensus coding scheme for discourse structure in dialogue was a non-trivial task. Discourse structure is many things to many researchers — attention, intentions, initiative, rhetorical structure, story trees, scripts, turn-taking behavior, etc. It was not feasible to devise and use a comprehensive coding scheme to cover all aspects of the discourse structure of dialogue. While there are already many existing taxonomies of discourse structure, none are completely satisfactory as general purpose coding schemes for dialogue. Many of the most thorough schemes have been devised for single-speaker text, and thus are problematic to apply directly to spontaneous dialogue. Many schemes devised for dialogue are appropriate only for certain genres of dialogue (e.g., classroom instruction), or for particular domains. Others are intended for radically different purposes than those of the computational linguistics dialogue community, e.g., focusing on some of the social relationships of the participants. Some of the range of taxonomies are described in [Traum, 1998], which proposes several dimensions by which to classify the type of structure. These dimensions include:

- Granularity: how much stuff (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 a short "sub-dialogue", exchange

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 types of units are allowed (one basic unit type, two or three types of units, or several types)?

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

Given the limited time and other resources available, it was not possible to attempt a comprehensive coding scheme for all aspects of discourse structure. Focusing coding of some content phenomena to only particular ranges of granularity allowed a principled way of restricting study to those aspects of the phenomena most central to the interests of the participants.

For our starting point we chose to focus on two 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 (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. In addition, these two schemes were linked by using the resulting structures from grounding analysis as basic input to the hierarchical intentional structures.

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.

Another issue is that dialogue participants can have intentions about many facets of dialogue, not all of which are connected or form coherent structures. Thus, one can have intentions about taking a turn, seizing the initiative, or augmenting the common ground, which will be only tenuously related (if at all) to the task-related intentions which the dialogue is about. For this reason, the dialogue act level of the DRI coding scheme [Carletta *et al.*, 1997a,

Allen and Core, Draft 1997] includes a field *information level* which concerns whether an utterance is doing the task, talking about the task, managing the communication, or other. To do a full-blown structural intentional analysis, one might need to maintain structures for each of these information levels, as well as hypotheses about how they might (are allowed to) inter-relate. Most of these non-task sorts of intentions occur at the meso-level or below. Thus restricting intentional analysis to macro-level grounded structures restricts the focus of analysis to task-related intentions, which are most compellingly argued to have a hierarchical structure.

Likewise, the phenomena of grounding, or adding information to the common ground, can occur on multiple levels. However, while some macro-level phenomena (such as the summarization and confirmation sub-dialogues which occur towards the ends of meetings and task oriented dialogues) definitely relate to maintenance of the common ground, they are very different in character from the local presentation and feedback phenomena (including acknowledgments and repairs) that characterize grounding at the meso-level. Thus restricting the grounding-relating coding to the meso-level, and treating the more macro-level confirmation and revision only insofar as it is a part of the intentional structure 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.[4] 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.

3.2.1 The Coding Scheme

The coding scheme used for pre-meeting coding exercises was distributed to the group members prior to coding assignments [Nakatani and Traum, 1998]. 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, 1998] 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 *common ground units* (CGUs), in which utterance tokens were gathered together in units of tokens which together served to add some material to the com-

[4]This hypothesis was proposed by Nakatani, personal communication, December 1997.

mon ground. Finally, the results of CGU coding was used as input for *Intentional/Informational Unit* (IU) Coding, ion which hierarchical intentional structure was built from either CGUs or smaller IUs. Each of these processes is briefly described in the subsections below.

Tokenization We felt that the segmentation guidelines from the previous meeting report ([Carletta *et al.*, 1997a], Section 5) were inappropriate to be used for discourse structure coding. Since those principles rely on the identification of dialogue act boundaries, they require such dialogue act identification to precede segmentation. While such act identification was the topic of the other group, we did not feel that the results had progressed sufficiently to require such act identification as a necessary pre-cursor to discourse structure analysis. Moreover, there is some feeling that meso-level grounding phenomena can occur within/below the level of illocutionary act performance. Finally, the reliability of such coding has not yet been demonstrated (κ below .4, although some high subclusters).

For these reasons, we instead chose to fall back on prosodic principles for dividing the dialogue into utterance tokens, using the intuition that a token should correspond to a single intonational phrase [Pierrehumbert, 1980].

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.

While much of the structure of CGUs corresponds to *initiative-response* pairs, as in the LINDA coding scheme [Dahlbäck and Jönsson, 1998], or dialogue games [Kowtko *et al.*, 1991, Carletta *et al.*, 1997b], there are some differences. These kinds of coding schemes attempt to encode *all* of the types of exchange behavior in dialogue, whereas CGUs are attempting to capture only those parts relating to mutual understanding.

As [Allwood *et al.*, 1992, Clark, 1994, Dillenbourg *et al.*, 1996] describe, there are multiple levels of coordination in dialogue. 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, which is not strictly a part of the grounding process. 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. Thus, a single token can be part of multiple CGUs. A good example is a suggestion followed by a refinement. 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.

The following principles summarize the decision procedures:

(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
3. **otherwise**, do not add this token to the CGU

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

Intentional Unit Analysis 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. Using 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 *supports* and *generates* [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 top-level node or nodes (i.e. nodes that are not dominated by any other node) are assigned identifiers 1...n, in order of linear occurrence. The children of any top-level node are identified as x.1 through x.n, where x is the number assigned to the dominating node and n is the total number of children. The next level nodes are assigned nodes x.y.1 through x.y.n, where x is the top-level dominating node, and y is the identifier of the immediately dominating node, and so on.

3.2.2 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, the idea being that results from the first set could influence details on the second set. While the close timing of the sets did not allow for careful analysis, comments from one of the participants did change the basis for the IU analysis in the second set to be based on a common CGU analysis (created by reconciling differences in the CGU coding of the co-chairs, which was performed before distribution of the assignments). The coding performed is summarized in table 2. The dialogues themselves, as well as some brief information about the corpora from which they were extracted and pointers to more information are included in the subsections below. IU analysis was not performed on the Maptask dialogue, since it was only a fragment, in which there was not much intentional structure present (the speakers were still in a phase of locating various objects on their maps). The Maptask and Verbmobil dialogues were also used for coding by the forward/backward dialogue act group.

TRAINS This 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

	Dialogue	CGU Coding	IU coding
1st exercise	TRAINS	11 coders	11 coders (own CGU)
	TOOT	11 coders	11 coders (own CGU)
2nd exercise	MAPTASK	9 coders	X
	Verbmobil	9 coders	9 coders (common CGU)

Table 2: Distribution of Pre-meeting coding exercises

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/`

The dialogue that was coded is shown starting with figure 3.

Toot 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, shown starting with figure 6, was provided by Diane Litman of AT&T Research.

Verbmobil 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.de/verbmobil/`

The dialogue we used for coding, r148c is shown starting with figure 9. In this dialogue, the two speakers try to establish a time and place for a meeting.

Maptask The DCIEM Map Task dialogs from which 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 provided, 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`. The portion of dialogue d204 used for the coding exercise can be found starting with figures 11 .

TRAINS Dialogue d93-8.2

```
Total Time: 2'30''      Total Turns: 66          Total Utterances : 93
S.1.1    hello
S.1.2    can I help you
U.2.1    how long does it take t- for a box t- car to get from
         Dansville to Corning
S.3.1    uh one hour
U.4.1    okay
U.4.2    I 'd like to take two boxcars
U.4.3    um and how long does it take to get a boxcar from
U.4.4    Bath to Corning
S.5.1    two hours
U.6.1    and how long does it take to load oranges
S.7.1    one hour
S.7.2    one hour per boxcar
U.8.1    per boxcar
U.9.1    and how long does it take to go from Corning to Dansville
S.10.1   no I 'm sorry
S.10.2   it takes one hour to load
S.10.3   any number of boxcars
S.10.4   I 'm sorry what was the se- next question
U.11.1   to go from Corning to Dansville with two boxcars of oranges
U.11.2   how does the long does that take
S.12.1   one hour
U.13.1   alright
U.13.2   I 'd like to have two boxcars leave Bath
S.14.1   mm-hm
U.15.1   pick up oranges in Corning
S.16.1   mm-hm
U.17.1   and move on to Dansville
S.18.1   okay you 'll need an engine
U.19.1   how many engines will I need for each of those
S.20.1   uh y(ou)- you just need one engine
S.20.2   to pull the boxcars
U.21.1   alright
U.21.2   so at midnight I 'd like for the can th- the boxcars
         need an engine to move anywhere
S.22.1   right
U.23.1   okay how long does it take to get an engine from Avon to Bath
S.24.1   four hours
```

Figure 3: TRAINS dialogue d93-8.2 p. 1

```
U.25.1   and to load the boxcars on
U.25.2   does that take time
S.26.1   no
S.26.2   uh to c- to couple to the boxcars is is instantaneous
U.27.1   okay
U.27.2   so if I were to move to Avon to Bath that would take me till four a.m.
S.28.1   right
U.29.1   from Bath to Corning will take me to six a.m.
S.30.1   right
U.31.1   loading the oranges will take me to seven a.m.
S.32.1   right
U.33.1   and moving to Dansville will take me to eight a.m.
S.34.1   that 's right
U.35.1   so I 'd like engine E one
U.35.2   to leave Avon
S.36.1   alright
U.37.1   go to Bath and pick up two boxcars
S.38.1   mm-hm
U.39.1   go to Corning and pick up oranges
S.40.1   mm-hm
U.41.1   and
U.41.2   then move on to Dansville
S.42.1   alright
U.43.1   okay
U.43.2   now at the same time
U.43.3   I 'd like for
U.43.4   two boxcars
U.43.5   to leave for an engine
U.43.6   how long does it take to get from Avon to Dansville
S.44.1   uh three hours
U.45.1   and how long does it take to load bananas
S.46.1   one hour
U.47.1   okay I 'd like for an engine to leave Avon
S.48.1   uh
S.48.2   the en- the engine from Avon is already leaving to go to Bath
U.49.1   did uh I use the
U.50.1   how long does it take from to get from Elmira to Corning
S.51.1   two hours
U.52.1   and from Corning to Dansville
S.53.1   one hour
```

Figure 4: TRAINS dialogue d93-8.2 p. 2

```
U.54.1   I 'd like engine E two
U.54.2   to go from Elmira to Corning
U.54.3   from Corning to Dansville
U.54.4   and pick up two boxcars
S.55.1   mm-hm
U.56.1   then go to da- from Dansville to Avon
S.57.1   alright
U.58.1   load the two boxcars with bananas
S.59.1   alright
U.60.1   and then move back to Dansville
S.61.1   alright
U.62.1   how long will that all take
S.63.1   ten hours
U.64.1   okay
U.64.2   I 'd like to do both of those things
S.65.1   alright
U.66.1   and they 're now finished
```

Figure 5: TRAINS dialogue d93-8.2 p. 3

Toot Dialogue

```
Total Time: 7'20''      Total Turns: 57         Total Utterances : 102
S.1.1    Hi
S.1.2    This is the AT&T Amtrak train schedule system
S.1.3    Which city do you want to go to?
U.2.1    Boston
S.3.1    What city do you want to go to Boston from?
U.4.1    Washington DC
S.5.1    On what day of the week do you want to go to Boston from Washington-D-C?
U.6.1    Thursday
U.6.2    3 thirty pm [ASR result: Thursday ]
S.7.1    What time do you want to leave on Thursday?
U.8.1    3 thirty pm
S.9.1    Do you want me to find the trains going to Boston from
         Washington-D-C on Thursday
S.9.2    around 3 thirty p-m now?
U.10.1   yes
S.11.1   I am going to query the Amtrak web site for you
S.11.2   This may takesome time if the internet traffic is heavy
S.11.3   Please wait patiently <SYSTEM PAUSE>
S.11.4   There are no trains going to Boston from Washington-D-C on Thursday
S.11.5   around 3 thirty p-m
S.11.6   The closest earlier train leaves at 2:15 pm every day
S.11.7   and it takes 9 hours
S.11.8   The closest later train leaves at 5:35 pm every day
S.11.9   and it takes 8 hours 50 minutes
S.11.10  Please say relax
S.11.11  to change your departure time or travel day
S.11.12  or say continue if my answer was sufficient
S.11.13  or =say re=peat to hear...
U.12.1   +relax+
S.13.1   Say "time" to relax your departure time
S.13.2   or say "date"
S.13.3   to relax your travel date
U.14.1   time
```

Figure 6: Toot dialogue p. 1

```
S.15.1   What time do you want to leave on Thursday ?
U.16.1   1 am
U.16.2   1 pm I'm sorry [ASR result: 1 a-m ]
S.17.1   Do you want me to find the trains going to Boston from
         Washington-D-C on Thursday
S.17.2   around 1 a-m now?
U.18.1   mm
U.18.2   cancel [ASR result: <reject> ]
S.19.1   I can't understand you
S.19.2   =Please re=peat
U.20.1   +no+
S.21.1   Do you want to continue and find a new set of trains?
U.22.1   yes
S.23.1   What city do you want to go to?
U.24.1   Boston
S.25.1   What city do you want to =go to Boston= from ?
U.26.1   +Washington DC+ [ASR result: Miami <due to barge-in overlap>]
S.27.1   On what day of the week do =you want to= go to Boston...
U.28.1   +Thursday+
S.29.1   What time do you want to leave on =Thursday= ?
U.30.1   +1 pm+
S.31.1   Do you want me to find the trains going to Boston from Miami on Thursday
S.31.2   around 1 p-m now?
U.32.1   no
S.33.1   Do you want to continue and find a new set of trains?
U.34.1   yes
S.35.1   What city do you want to go to?
U.36.1   Boston
S.37.1   What city do you want to go to Boston =from= ?
U.38.1   +Washing+ton DC
S.39.1   On what day of the week do you want to go to Boston from Washington-D-C ?
U.40.1   Thursday
S.41.1   What time do you =want to leave on Thurs=day ?
U.42.1   +3 thirty pm+
S.43.1   Do you want me to find the trains going to Boston from
         Washington-D-C on Thursday
S.43.2   around 3 thirty p=m= now?
U.44.1   +yes+
```

Figure 7: Toot dialogue p. 2

```
S.45.1   There are no trains going to Boston from Washington-D-C on Thursday
S.45.2   around 3 thirty p-m
S.45.3   The closest earlier train leaves at 2:15 pm every day
S.45.4   and it takes 9 hours
S.45.5   The closest later train leaves at 5:35 pm every day
S.45.6   and it takes 8 hours 50 minutes
S.45.7   Please say relax
S.45.8   to change your departure time or travel day
S.45.9   or say con=tinue= if my answer was...
U.46.1   +relax+
S.47.1   Say "time" to relax your =depar=ture time
U.48.1   +time+
S.49.1   What time do you want to leave on Thursday ?
U.50.1   1 p-m
S.51.1   Do you want me to find the trains going to Boston from
         Washington-D-C on Thursday
S.51.2   around 1 p-m now?
U.52.1   yes
S.53.1   I am going to query the Amtrak web site for you
S.53.2   This may take some time if the internet traffic is heavy
S.53.3   Please wait patiently <SYSTEM PAUSE>
S.53.4   There are 2 trains available around 1 p-m
S.53.5   The first train leaves at 12:10 pm every day
S.53.6   and it takes 10 hours 10 minutes
S.53.7   The second train leaves at 1:10 pm from Sunday to Friday
S.53.8   and it takes 8 hours 10 minutes
S.53.9   Do you want to continue and find a new set of trains?"
U.54.1   no [ASR result: <reject> ]
S.55.1   I didn't hear anything from you
S.55.2   Say "yes" to find a new set of trains
S.55.3   or say "no" to end the dialogue
U.56.1   no
U.56.2   I'm done here
S.57.1   Thank you for using AT&T Amtrak schedule system
S.57.2   See you next time
```

Figure 8: Toot dialogue p. 3

Verbmobil Dialogue r148c

```
Total Time: ??         Total Turns: 20          Total Utterances : 53
A.1.1  so
A.1.2  we should meet again
A.1.3  how 'bout
A.1.4   uh
A.1.5  how 'bout next week
A.1.6  what day are good for you
A.1.7  what days are good for you
B.2.1  actually next week I am on vacation
A.3.1  gosh
A.3.2  I guess we will have to meet the week after that
A.3.3  umm,
A.3.4  how 'bout Monday
B.4.1  uhh Monday the tenth
A.5.1  uh-huh
B.6.1  well unfortunately my vacation uh runs through the fourteenth
B.6.2  and I have nonrefundable plane tickets
B.6.3  I was planning on being on a beach in Acapulco about that point
A.7.1  well
A.7.2  when are you getting back
B.8.1  I get back on the fifteenth
B.8.2  rest up on the sixteenth
B.8.3  which is a Sunday
B.8.4  and I am back at work on the seventeenth
B.8.5  but I have a seminar all day
B.8.6  I think the first day that's really good for me
B.8.7  is the eighteenth
B.8.8  that's a Tuesday
A.9.1  okay
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
```

Figure 9: Verbmobil dialogue r148c p. 1

```
B.14.1  I say if I have got enough money to go to Acapulco
B.14.2  I have got enough money to go to one of those silly places on Craig street
B.14.3  how about Great Scott
A.15.1  sounds great except they have been out of business for a while
A.15.2  how about some other place
A.15.3  let us just wander up Craig
A.15.4  and pick one we like that day
B.16.1  that sounds pretty good
B.16.2  okay
B.16.3  umm
B.16.4  I will meet you outside Cyert Hall
B.16.5  at noon
B.16.6  does that sound alright for you
A.17.1  see you then
B.18.1  roger over and out
A.19.1  thought it was roger wilco
B.20.1  oh no it is what we always say when we are talking on screen
```

Figure 10: Verbmobil dialogue r148c p. 2

Maptask Dialogue d204

```
Total Time: ??         Total Turns: 34         Total Utterances : 61
G.1.1  Okay
G.1.2  Do you have a start X
F.2.1  =Yeah=
G.3.1  +up top+
F.4.1  +Up+ by sandy shore
G.5.1  By sandy shore
G.5.2  Just below that do you have a well
F.6.1  No
G.7.1  Oh
G.7.2  Great
G.7.3  Do you have any land marks just below
F.8.1  Uh no
F.8.2  About half about a quar- th-
F.8.3  third of the way down I have some hills
G.9.1  Okay
F.10.1  And I got a uh
F.10.2  local resident
F.10.3  and an iron bridge
G.11.1  An iron bridge
F.12.1  In the middle of the map
F.12.2  right at the top
G.13.1  Okay
G.13.2  I don't have the iron bridge
F.14.1  Do you got the =my=
G.15.1  +Do+ y-
F.16.1  You got the local resident
G.17.1  Yes I do
F.18.1  Okay, my iron bridge is right above that
G.19.1  =Oh, okay=
F.20.1  +like it's+ like uh
F.20.2  =going across=
G.21.1  +It crosses+
G.21.2  =the bay=
F.22.1  +go-+
F.22.2  Yeah
```

Figure 11: Maptask dialogue d203 p. 1

```
F.22.3  Going across the the river there
F.22.4  Like
F.22.5  below the local residents
F.22.6  about an inch
F.22.7  to the left
G.23.1  Is
G.23.2  the hills
F.24.1  I got the h-
F.24.2  That's where my hills are
G.25.1  Okay
F.26.1  Um to the
F.26.2  if you go right of the iron bridge
F.26.3  I have a woodland
G.27.1  See Okay
F.28.1  =Okay=
G.29.1  +I have a+
F.30.1  if you follow
F.30.2  if you follow the brook down
F.30.3  =the babbling brook=
G.31.1  +The forked stream+
F.32.1  The babbling brook
F.32.2  or whatever
F.32.3  I have a dead tree
F.32.4  Do you have that
G.33.1  I have the dead tree at the fork
F.34.1  Yeah
```

Figure 12: Maptask dialogue d203 p. 2

```
1      "suggest meet, ask days"
    A.1.2  we should meet again
    A.1.3  how 'bout
    A.1.5  how 'bout next week
    A.1.6  what day are good for you
    A.1.7  what days are good for you
    B.2.1  actually next week I am on vacation
2      "on vacation next week"
    B.2.1  actually next week I am on vacation
    A.3.1  gosh
3      "ask following monday"
    A.3.2  I guess we will have to meet the week after that
    A.3.4  how 'bout Monday
    B.4.1  uhh Monday the tenth
    A.5.1  uh-huh
4      "reject"
    B.6.1  well unfortunately my vacation uh runs through
           the fourteenth
    B.6.2  and I have nonrefundable plane tickets
    B.6.3  I was planning on being on a beach in Acapulco about
           that point
    A.7.1  well
5      "ask when back?"
    A.7.2  when are you getting back
    B.8.1  I get back on the fifteenth
6      "tuesday 18th is first opportunity"
    B.8.1  I get back on the fifteenth
    B.8.2  rest up on the sixteenth
    B.8.3  which is a Sunday
    B.8.4  and I am back at work on the seventeenth
    B.8.5  but I have a seminar all day
    B.8.6  I think the first day that's really good for me
    B.8.7  is the eighteenth
    B.8.8  that's a Tuesday
    A.9.1  okay
```

Figure 13: Verbmobil CGU coding p. 1

3.2.3 Sample Codings for Verbmobil Dialogue

In this section we present sample codings for CGU and IU analysis. The next sections give further details on agreement and consensus codings among the 8-12 codings of each dialogue. As stated in section 3.2.1, coding proceeds in three phases. The first phase, tokenization, is shown in Section 3.2.2. This was used as input for CGU coding. In Figures 13 and 14, we present the unified coding from Nakatani and Traum, which was used as common input input for the IU coding of this dialogue.

Figure 15 shows a sample IU coding using the CGU coding in Figures 13,14 as input.

```
7      "suggest lunch"
    A.9.2  want to have lunch
    B.10.1  that sounds pretty good
8      "accept lunch, suggest before noon"
    B.10.2  are you available just before noon
    A.11.1  we can meet at noon
9      "meet at noon"
    A.11.1  we can meet at noon
    B.12.1  sounds good
10     "ask on campus?"
    B.12.3  on campus or off
    A.13.1  your choice
11     "your choice"
    A.13.1  your choice
    B.14.1  I say if I have got enough money to go to Acapulco
12     "off campus (implicit) suggest Craig st, Great scott"
    B.14.1  I say if I have got enough money to go to Acapulco
    B.14.2  I have got enough money to go to one of those silly places on Craig street
    B.14.3  how about Great Scott
    A.15.1  sounds great except they have been out of business for a while
13     "reject Scott, reason, accept Craig, suggest defer decision"
    A.15.1  sounds great except they have been out of business for a while
    A.15.2  how about some other place
    A.15.3  let us just wander up Craig
    A.15.4  and pick one we like that day
    B.16.1  that sounds pretty good
14     "set meeting place"
    B.16.4  I will meet you outside Cyert Hall
    B.16.5  at noon
    B.16.6  does that sound alright for you
    A.17.1  see you then
15     "goodbye"
    B.18.1  roger over and out
    A.19.1  thought it was roger wilco
    B.20.1  oh no it is what we always say when we are talking on screen
```

Figure 14: Verbmobil CGU coding p. 2

```
iu.1    "plan to meet (again)"
        iu.1.1
                "select meeting day"
        1       A.1.2, A.1.3, A.1.5, A.1.6, A.1.7, B.2.1
                                "suggest meet, ask days"
        2       B.2.1, A.3.1
                                "on vacation next week"
        3       A.3.2, A.3.4, B.4.1, A.5.1
                                "ask following monday"
        4       B.6.1, B.6.2, B.6.3, A.7.1
                                "reject"
        5       A.7.2, B.8.1
                                "ask when back?"
        6       B.8.1, B.8.2, B.8.3, B.8.4, B.8.5, B.8.6, B.8.7,
                B.8.8, A.9.1
                                "tuesday 18th is first opportunity"
        iu.1.2
                "set meeting time"
        7       A.9.2, B.10.1
                                 "suggest lunch"
        8       B.10.2, A.11.1
                                "accept lunch, suggest before noon"
        9       A.11.1, B.12.1
                                "meet at noon"
        iu.1.3
                "select place for lunch"
        10      B.12.3, A.13.1
                                "ask on campus?"
        11      A.13.1, B.14.1
                                "your choice"
        12      B.14.1, B.14.2, B.14.3, A.15.1
                               "off campus (implicit) suggest Craig st,
                                Great scott"
        13      A.15.1, A.15.2, A.15.3, A.15.4, B.16.1
                                "reject Scott, reason, accept Craig,
                                 suggest defer decision"
        iu.1.4
                "set place to meet"
        14      B.16.4, B.16.5, B.16.6, A.17.1
                                "set meeting place"
iu.2    "closing"
        15      B.18.1, A.19.1, B.20.1
                                "goodbye"
```

Figure 15: Verbmobil IU coding

3.2.4 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.

Majority Codings Below are shown the best attempt at an induced majority CGU coding, by taking the units with the highest agreement, and occasionally filling in with the most consistent matches in case of divergences. A few places in the dialogue had too much disagreement to be able to select a particular way of coding that stretch. These are indicated with markings `vvvvvvvvvvvvv` above and `^^^^^^^^^^^^^` below, with all coded units in between. The original utterance tokens that comprise the units are shown in the dialogue transcripts in Section 3.2.2. Also shown are the number of coders who included that unit, and the coders (anonymized to a single letter) and the position in the dialogue the unit occurred (with 1 being the first CGU marked, etc.)

TRAINS majority Coding

```
ratio: 5/11   unit: S.1.1,S.1.2,U.2.1
ratio: 9/11   unit: U.2.1,S.3.1
ratio: 9/11   unit: S.3.1,U.4.1
ratio: 5/11   unit: U.4.2
ratio: 8/11   unit: U.4.3,U.4.4,S.5.1
ratio: 6/11   unit: S.5.1,U.6.1
ratio: 7/11   unit: U.6.1,S.7.1
ratio: 2/11   unit: S.7.1,S.7.2,U.8.1,S.10.1,S.10.2,S.10.3
ratio: 6/11   unit: U.9.1,S.10.4,U.11.1,U.11.2,S.12.1
ratio: 9/11   unit: S.12.1,U.13.1
ratio: all   unit: U.13.2,S.14.1
ratio: all   unit: U.15.1,S.16.1
ratio: all   unit: U.17.1,S.18.1
ratio: 8/11   unit: S.18.1,U.19.1
ratio: 6/11   unit: U.19.1,S.20.1
ratio: 7/11   unit: S.20.1,S.20.2,U.21.1
ratio: 8/11   unit: U.21.2,S.22.1
ratio: all   unit: U.23.1,S.24.1
ratio: 5/11   unit: S.24.1,U.25.1
ratio: 7/11   unit: U.25.1,U.25.2,S.26.1
ratio: 8/11   unit: S.26.1,S.26.2,U.27.1
ratio: all   unit: U.27.2,S.28.1
ratio: all   unit: U.29.1,S.30.1
ratio: all   unit: U.31.1,S.32.1
ratio: all   unit: U.33.1,S.34.1
ratio: all   unit: U.35.1,U.35.2,S.36.1
ratio: all   unit: U.37.1,S.38.1
```

```
ratio: all   unit: U.39.1,S.40.1
ratio: 8/11   unit: U.41.1,U.41.2,S.42.1
ratio: 5/11   unit: U.43.6,S.44.1
ratio: 6/11   unit: S.44.1,U.45.1
ratio: all   unit: U.45.1,S.46.1
ratio: 6/11   unit: S.46.1,U.47.1

vvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvv
ratio: 1/11   unit: U.47.1,S.48.1,S.48.2,U.49.1
ratio: 2/11   unit: U.47.1,S.48.2
ratio: 3/11   unit: U.47.1,S.48.1
ratio: 2/11   unit: U.47.1,S.48.1,S.48.2
ratio: 1/11   unit: U.47.1,S.48.2,U.49.1
ratio: 1/11   unit: U.47.1,S.48.1,U.49.1
ratio: 1/11   unit: S.48.1,S.48.2,U.50.1
ratio: 4/11   unit: S.48.2,U.49.1
ratio: 1/11   unit: S.48.2,U.50.1
ratio: 1/11   unit: U.49.1
^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
ratio: all   unit: U.50.1,S.51.1
ratio: 6/11   unit: S.51.1,U.52.1
ratio: all   unit: U.52.1,S.53.1
ratio: 6/11   unit: S.53.1,U.54.1
ratio: all   unit: U.54.1,U.54.2,U.54.3,U.54.4,S.55.1
ratio: all   unit: U.56.1,S.57.1
ratio: all   unit: U.58.1,S.59.1
ratio: all   unit: U.60.1,S.61.1
ratio: 10/11   unit: U.62.1,S.63.1
ratio: 8/11   unit: S.63.1,U.64.1
ratio: all   unit: U.64.2,S.65.1
ratio: 9/11   unit: U.66.1
```

TOOT majority Coding

```
ratio: 6/11   unit: S.1.1,S.1.2,S.1.3,U.2.1
ratio: 9/11   unit: U.2.1,S.3.1
ratio: all   unit: S.3.1,U.4.1
ratio: 9/11   unit: U.4.1,S.5.1
ratio: 9/11   unit: S.5.1,U.6.1
ratio: 3/11   unit: U.6.1,U.6.2,S.7.1
ratio: 8/11   unit: S.7.1,U.8.1
ratio: 4/11   unit: U.8.1,S.9.1
ratio: all   unit: S.9.1,S.9.2,U.10.1
ratio: 7/11   unit: U.10.1,S.11.1
ratio: 5/11   unit: S.11.1,S.11.2,S.11.3
ratio: 3/11   unit: S.11.4,S.11.5,S.11.6,S.11.7,S.11.8,S.11.9,
                        S.11.10,S.11.11,S.11.12,S.11.13,U.12.1
ratio: 7/11   unit: U.12.1,S.13.1
ratio: all   unit: S.13.1,S.13.2,S.13.3,U.14.1
ratio: 8/11   unit: U.14.1,S.15.1
ratio: 6/11   unit: S.15.1,U.16.1

vvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvv
```

```
ratio: 2/11   unit: U.16.1,U.16.2,S.17.1
ratio: 1/11   unit: U.16.1,U.16.2,S.17.1,S.17.2,U.18.2,S.19.1,
                    S.19.2,U.20.1,S.21.1
ratio: 1/11   unit: U.16.1,S.17.1,S.17.2,U.18.1,U.18.2,S.19.1,
                    S.19.2,U.20.1
ratio: 2/11   unit: S.17.1,S.17.2,U.18.2
ratio: 3/11   unit: S.17.1,S.17.2,U.18.1
ratio: 1/11   unit: S.17.1,S.17.2,U.18.1,U.18.2,S.19.1,S.19.2,U.20.1
ratio: 1/11   unit: S.17.1,S.17.2,U.20.1
ratio: 1/11   unit: S.17.1,S.17.2,U.18.1,U.18.2
ratio: 1/11   unit: U.18.1,U.18.2,S.19.1
ratio: 2/11   unit: U.18.2,S.19.1,S.19.2,U.20.1
ratio: 1/11   unit: U.18.2,S.19.1,S.19.2,U.20.1,S.21.1
ratio: 1/11   unit: U.18.2
ratio: 2/11   unit: S.19.1,S.19.2,U.20.1
ratio: 1/11   unit: S.19.1,S.19.2.,U.20.1
ratio: 4/11   unit: U.20.1,S.21.1
^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^

ratio: all   unit: S.21.1,U.22.1
ratio: 7/11   unit: U.22.1,S.23.1
ratio: all   unit: S.23.1,U.24.1
ratio: 9/11   unit: U.24.1,S.25.1
ratio: all   unit: S.25.1,U.26.1
ratio: 7/11   unit: U.26.1,S.27.1
ratio: 10/11   unit: S.27.1,U.28.1
ratio: 8/11   unit: U.28.1,S.29.1
ratio: 10/11   unit: S.29.1,U.30.1
ratio: 4/11   unit: U.30.1,S.31.1
ratio: 9/11   unit: S.31.1,S.31.2,U.32.1
ratio: 7/11   unit: U.32.1,S.33.1
ratio: all   unit: S.33.1,U.34.1
ratio: 8/11   unit: U.34.1,S.35.1
ratio: all   unit: S.35.1,U.36.1
ratio: 9/11   unit: U.36.1,S.37.1
ratio: all   unit: S.37.1,U.38.1
ratio: 9/11   unit: U.38.1,S.39.1
ratio: all   unit: S.39.1,U.40.1
ratio: 9/11   unit: U.40.1,S.41.1
ratio: all   unit: S.41.1,U.42.1
ratio: 6/11   unit: U.42.1,S.43.1
ratio: all   unit: S.43.1,S.43.2,U.44.1
ratio: 4/11   unit: U.44.1,S.45.1
ratio: 5/11   unit: S.45.1,S.45.2,S.45.3,S.45.4,S.45.5,S.45.6,S.45.7,
                    S.45.8,S.45.9,U.46.1
ratio: 8/11   unit: U.46.1,S.47.1
ratio: 10/11   unit: S.47.1,U.48.1
ratio: 8/11   unit: U.48.1,S.49.1
ratio: all   unit: S.49.1,U.50.1
ratio: 4/11   unit: U.50.1,S.51.1,S.51.2
ratio: all   unit: S.51.1,S.51.2,U.52.1
ratio: 8/11   unit: U.52.1,S.53.1
ratio: 7/11   unit: S.53.1,S.53.2,S.53.3
```

```
vvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvv
ratio: 2/11   unit: S.53.1,S.53.2,S.53.3,S.53.4,S.53.5,S.53.6,S.53.7,
                        S.53.8,S.53.9,U.54.1
ratio: 1/11   unit: S.53.1,S.53.2,S.53.3,S.53.4,S.53.5,S.53.6,S.53.7,
                        S.53.8,U.54.1
ratio: 1/11   unit: S.53.1,S.53.2,S.53.3,S.53.4,S.53.5,S.53.6,S.53.7,
                        S.53.8,S.53.9,U.54.1,S.55.1,S.55.2,S.55.3,U.56.1
ratio: 2/11   unit: S.53.4,S.53.5,S.53.6,S.53.7,S.53.8,S.53.9,U.54.1
ratio: 3/11   unit: S.53.4,S.53.5,S.53.6,S.53.7,S.53.8
ratio: 2/11   unit: S.53.4,S.53.5,S.53.6,S.53.7,S.53.8,S.53.9,U.54.1,
                        S.55.1,S.55.2,S.55.3,U.56.1
ratio: 1/11   unit: S.53.9,U.54.1,S.55.1,S.55.2,S.55.3,U.56.1
ratio: 1/11   unit: S.53.9,U.54.1
ratio: 1/11   unit: S.53.9,U.54.1,S.55.1,S.55.2,S.55.3,U.56.1,U.56.2
ratio: 1/11   unit: S.53.9,U.56.1
ratio: 1/11   unit: U.54.1,S.55.1,S.55.2,S.55.3,U.56.1
ratio: 2/11   unit: U.54.1,S.55.1,S.55.2,S.55.3,U.56.1,U.56.2,S.57.1
ratio: 1/11   unit: U.54.1
ratio: 2/11   unit: S.55.1,S.55.2,S.55.3,U.56.1
ratio: 1/11   unit: U.56.1,U.56.2
ratio: 3/11   unit: U.56.1,U.56.2,S.57.1
ratio: 1/11   unit: U.56.1,S.57.1
ratio: 1/11   unit: U.56.1,U.56.2,S.57.1,S.57.2
ratio: 1/11   unit: U.56.2,S.57.1,S.57.2
ratio: 1/11   unit: U.56.2
ratio: 4/11   unit: S.57.1,S.57.2
^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
```

Verbmobil majority Coding

```
ratio: 2/8   unit: A.1.2,A.1.3,A.1.5,A.1.6,A.1.7,B.2.1
ratio: 7/8   unit: B.2.1,A.3.1
ratio: 4/8   unit: A.3.2,A.3.4,B.4.1
ratio: 5/8   unit: B.4.1,A.5.1
ratio: 7/8   unit: B.6.1,B.6.2,B.6.3,A.7.1
ratio: 6/8   unit: A.7.2,B.8.1
ratio: 6/8   unit: B.8.1,B.8.2,B.8.3,B.8.4,B.8.5,B.8.6,B.8.7,
                        B.8.8,A.9.1
ratio: all   unit: A.9.2,B.10.1
ratio: 6/8   unit: B.10.2,A.11.1
ratio: all   unit: A.11.1,B.12.1
ratio: 6/8   unit: B.12.3,A.13.1
ratio: 5/8   unit: A.13.1,B.14.1
ratio: 6/8   unit: B.14.1,B.14.2,B.14.3,A.15.1
ratio: 6/8   unit: A.15.1,A.15.2,A.15.3,A.15.4,B.16.1
ratio: 6/8   unit: B.16.4,B.16.5,B.16.6,A.17.1
ratio: 4/8   unit: B.18.1,A.19.1,B.20.1
```

Maptask majority Coding

```
ratio: 4/9   unit: G.1.2,F.2.1
ratio: 4/9   unit: F.4.1,G.5.1
```

```
ratio: 8/9   unit: G.5.2,F.6.1
ratio: 4/9   unit: F.6.1,G.7.1
ratio: all   unit: G.7.3,F.8.1
ratio: 5/9   unit: F.8.2,F.8.3,G.9.1
ratio: 5/9   unit: F.10.1,F.10.2,F.10.3,G.11.1
ratio: 6/9   unit: F.12.1,F.12.2,G.13.1
ratio: 6/9   unit: G.13.2,F.14.1
ratio: 3/9   unit: G.15.1
ratio: 5/9   unit: F.16.1,G.17.1
ratio: 6/9  unit: G.17.1,F.18.1
ratio: 7/9   unit: F.18.1,G.19.1

vvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvv
ratio: 1/9   unit: F.18.1,G.19.1,F.20.1,F.20.2,G.21.1,G.21.2,F.22.1,
                       F.22.2,F.22.3,F.22.4,F.22.5
ratio: 1/9   unit: F.18.1,G.19.1,F.20.1,F.20.2,G.21.1,G.21.2,F.22.1,
                       F.22.2,F.22.3
ratio: 1/9   unit: G.19.1,F.20.1,F.20.2,G.21.1,G.21.2,F.22.1,F.22.2
ratio: 1/9   unit: F.20.1,F.20.2,G.21.1,G.21.2,F.22.1,F.22.2
ratio: 1/9   unit: F.20.1,F.20.2,F.22.1
ratio: 1/9   unit: F.20.1,F.20.2,G.21.1
ratio: 1/9   unit: F.20.1,F.20.2,G.21.1,G.21.2
ratio: 1/9   unit: F.20.2,G.21.1,G.21.2,F.22.1,F.22.2,F.22.3
ratio: 2/9   unit: G.21.1,G.21.2,F.22.2
ratio: 1/9   unit: G.21.1,G.21.2,F.22.2,F.22.3
ratio: 3/9   unit: F.22.3,F.22.4,F.22.5,F.22.6,F.22.7,G.23.1
ratio: 2/9   unit: F.22.5,F.22.6,F.22.7,G.23.1,G.23.2,F.24.1,F.24.2,G.25.1
ratio: 1/9   unit: F.22.6,F.22.7,G.23.1,G.23.2,F.24.1,F.24.2,G.25.1
ratio: 2/9   unit: G.23.1,G.23.2,F.24.1,F.24.2,G.25.1
ratio: 1/9   unit: G.23.1,G.23.2,F.24.1
ratio: 1/9   unit: G.23.1,G.23.2
ratio: 1/9   unit: G.23.1,G.23.2,F.24.1,F.24.2
ratio: 1/9   unit: G.23.2,F.24.1
^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^

ratio: 3/9   unit: F.24.1,F.24.2,G.25.1
ratio: 5/9   unit: F.26.1,F.26.2,F.26.3,G.27.1

vvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvv
ratio: 2/9   unit: G.29.1
ratio: 2/9   unit: F.30.1,F.30.2,F.30.3,G.31.1,F.32.1,F.32.2,F.32.3,
                       F.32.4,G.33.1
ratio: 2/9   unit: F.30.1,F.30.2,F.30.3,G.31.1
ratio: 1/9   unit: F.30.1,F.30.2,G.31.1
ratio: 1/9   unit: F.30.1,F.30.2,F.30.3,G.31.1,F.32.1,F.32.2,F.32.3,
                       F.32.4,G.33.1,F.34.1
ratio: 1/9   unit: F.30.1,F.30.2,F.30.3
ratio: 1/9   unit: F.30.2,F.30.3,G.31.1,F.32.1,F.32.2,G.33.1
ratio: 1/9   unit: F.30.2,F.32.3
ratio: 1/9   unit: F.30.3
ratio: 2/9   unit: G.31.1,F.32.1,F.32.2
ratio: 1/9   unit: G.31.1,F.32.2
ratio: 1/9   unit: G.31.1,F.32.1
```

```
ratio: 1/9   unit: F.32.1,F.32.2,F.32.3,F.32.4,G.33.1
ratio: 1/9   unit: F.32.1,F.32.2,G.33.1
^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
ratio: 3/9   unit: F.32.3,F.32.4,G.33.1
ratio: 7/9    unit: G.33.1,F.34.1
```

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 3 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.

Dialogue	**avg %**
TRAINS	0.41
TOOT	0.36
Verbmobil	0.30
MAPTASK	0.26

Table 3: Average coders per proposed CGU

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.

As an example, if someone coded a dialogue consisting of the following utterance tokens:

```
1.1
1.2
2.1
2.2
2.3
3.1
```

Into the following CGUs:

```
1       1.1, 1.2, 2.1
2       2.1, 2.3, 3.1
```

The following "acts" would be assigned:

```
 begin    middle   end
1.1     1       0       0
1.2     0       1       0
2.1     1       0       1
2.2     0       0       0
2.3     0       1       0
3.1     0       0       1
```

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*).

	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 4: CGU Inter-coder reliability

From this, it was possible to compute agreement and expected agreement (by examining the relative frequencies of these tags), and thus κ [Siegel and Castellan, 1988]. The numbers for the group as a whole are shown in table 4 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.

Difficult Issues in CGU Coding The following qualitative problem areas and phenomena were noted from analyzing the sources of disagreement in the codings of these dialogues:

Verbmobil:

1. restart-continue: in or out? Manual says borderline. What about issue of content? specific issue: A.1.3, A.1.6

2. What to do about additional acknowledgments, which have a different style or provide better evidence - add to prior CGU? case in point a.3.2 , b.16.2

3. request repair vs clarification vs repair
   unit 3 - two coders treated 4.1 as repair request, 5.1 as repair, and 6.1 as ack, others treated 4.1 as repair and 5.1 as ack, while 4 coders treated 4.1 as ack and new initiate (one coder also did the latter, to encode the sub-dialogue grounding), if so, does 5.1 need to be an init of answer?

4. inclusion of irrelevant material in cgu that is not directly acked - e.g., 6.2,6.3, also 8.2-8.5,14.1-14.2

5. starting cgus from acceptances.

6. oddness at beginning closing of dialogue, due perhaps to strange recording situation, leads to disagreements.

TRAINS:

1. repair around 7.2 - does it go with 7.1, and if so, with 6.1 also? or 9.1? also overlap, 8.1,7.2 unit from 2 people, transcript problems with trains -continuing different treatments of turns 9,10,11.

2. 19.1, beginning of repair sequence or simple ack with followup question?

3. multi-utterance answer to question - all or just first as ack (manual says first, but cf. 20.2 responses), also 26.1, 26.2, answer with elaboration

4. 24.1 - can't leave off answer the same way as acceptance. But "and" does seem to ack, or could just be continuation - maybe prosody helps?

5. which things don't need to initiate a unit. E.g., various kinds of positive feedback of agreement or yes answer to check question? Also e.g., 44.1 some contentful answers not seen as initiating, or not seen as implicitly grounded. Several other places, same pattern for Q&A (6.1.4)

6. turn 43 controversy over cancelations, how to mark them well for agreement - basically some coders agree that none of those utterances enter common ground, but marked three diff. ways

7. big confusion around turn 47 - 49: overlap? some people identify the same units, but * or don't *. Too tough to come up with consensus.

8. look at relation to backward understanding acts, notion of keeping cgu's open - useful at all for speech? use cleaned up speech (speech repairs and disfluencies cleaned up) for CGU coding

9. Q&A in same CGU if answer *complex*? -difficult to tell about discourse marker vs. ack

10. 13.2 continues content of 4.2, also role of 9.1, 11.1

Maptask:

1. Big Q: did everyone listen carefully to speech - intonation and timing very important here.

2. overlap 2.1 and 3.1: 3.1 seems to be a repair given lack of immediate response to 1, thus whole thing to 5 seems one cgu; similar phenomenon in 20.1 coming at same time as 19.1. (different subcodings, but all see endings at 5.1, then agree about 5.2)

3. how to do 7, with ack then acceptance - include both or only one?
4. 8.1 in next unit? complex answer and explanation, so seems yes), repair which is not complete restart in 8.2, also 13.2 is negative response after ack of 13.1)
5. 11.1 seems like reqrepair not ack, hence joined with next unit
6. 14.1 as implicit ack of 13.1, also, with 16.1 followup of 10.2, and by implication ack of 13.1, since asking about not-mentioned part.
7. 16.1 unit should include 14.1, because not complete restart - need q part (perhaps clear from intonation).
8. 20.1 as repair when ack comes too late.
9. complete disagreement about where to include 20.2! 8 coders and 8 different codings for this unit
10. transcription error in 22.4, really just disfluency
11. multiple ack? 28.1, included in previous unit or initiating unit?
12. very different codings for 30.1-32.2

TOOT:

1. starting new cgu on different topic without ack, e.g., 1.1,1.2 from 1.3
2. 6.2, extra info, misunderstood - combine?
3. question about splitting 11 at 11.3 or 11.0 or not at all.
4. repair also counting for ack? 16.2 - don't think it should)
5. general confusion from 16.1 to 20.1
6. initiate unit for yes answer 22.1
7. missing bits in transcript in turn 45 - some tried to add uu tokens
8. new unit at 45.7?
9. turn 51 ack, 51.1 or 51.2 or both - why split there?

3.2.5 IU Coding Analysis

By Christine Nakatani

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 uniform IUs for all the coders. Thus, the reliability for IU coding could

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

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 κ 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 κ score.

Table 5 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.

Table 6 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), κ scores can be computed. Table 7 shows the κ scores measuring the reliability of the codings for each pair of labelers.

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

By examining Table 5, 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-intiial CGUs. A weaker reliability metric on the pooled data from nine coders,

CODERID	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 6: Observed agreement for IU-initial CGUs

CODERID	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 7: Pairwise κ scores

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 κ score is .80, which is highly reliable.

3.3 Meeting Summary

3.3.1 Participants

Ellen Gurman Bard	University of Edinburgh
Jennifer Chu-Carroll	Bell Labs.
Peter Heeman	Oregon Graduate Institue of Science and Technology
Julia Hirschberg	AT&T Labs.
Koiti Hasida	ElectroTechnical Lab.
Yasuo Horiuchi	Chiba University
Yasuhiro Katagiri	ATR Media Integration and Comm. Research Labs.
Diane Litman	AT&T Labs.
Kikuo Maekawa	National Language Research Institute
Christine H. Nakatani	Bell Labs.
Owen Rambow	CoGenTex
Michael Strube	University of Pennsylvania
Masafumi Tamoto	NTT Basic Research Labs.
Yuka Tateishi	University of Tokyo
David Traum	University of Maryland
Jennifer J. Venditti	Ohio State University
Gregory Ward	Northwestern University

3.3.2 Session Reports

Day 1: Report by Peter Heeman

Roughly half of the participants of the discourse structure group had done the homework assignments of coding grounding units and using these grounding units as the basis for coding intentional structure. Linking intentional structure to grounding units is a new proposal, and considerable time was spent in the first day evaluating the worthiness of this proposal, in particular coding grounding units.

Issue 1: Why do we even want to do CGUs?

We need to motivate why people should want to code grounding units. What can people do with such an annotation? We need to remember that motivations should not just concern those working on the scheme.

David Traum presented an overview of a paper that he is working on [Traum, 1998], in which he surveys existing annotation schemes in terms of granularity of units, content and relationship between units. The result of this is classifying annotation schemes by what they code at the micro, meso and

macro level (see Section 3.2). There seems to be general consensus that discourse structure is real, and hence if CGUs can simplify IU analysis, that could be motivation in itself. CGUs simplify IU coding since they factor how all of the conversation concerned with grounding, such as backchannels, paraphrases, repetitions, confirmations, (some) clarification subdialogues, as well as even some turn-taking and initiative issues. This allows people interested in intentional structure, but not in grounding, to factor out the grounding issues when doing intentional structure.

Coding grounding is also important in itself. Spoken dialogue systems will need to deal with misrecognitions and misunderstandings, and that is what coding grounding is partially about. Having an annotated corpus will allows us to study how humans accomplish this, thus allowing us to build more natural systems. For instance, we might want an animated agent to give head nods for acknowledgments. It was noted that work by Herbert Clark (e.g., [Clark and Marshall, 1981]) showed that participants change description once referents enter common ground.

But, as some group members noted, head nods, backchannels can come at any point, not necessarily at certain points. How small are CGUs?

However, there was not complete agreement. CGUs are just one choice for doing meso level analysis. Other possibilities are adjacency pairs and turns. Furthermore, more thought needs to be given to using CGUs as a basic unit for doing IU analysis. CGUs buys us into grounding theory, whereas using intonational phrases, say, would be more theory neutral. One proposal was to even use rhetorical relations (or some similar theory) for the whole structure.

Given that we do code CGUs, there is still the issue of how well these can be used as the basic unit for IU analysis. A CGU might involve two different intentions, which need to go into two different IUs. Part of the problem here is confusion with the role of the speech that provides evidence that prior speech was grounded, such as a relevant next turn. This kind of utterance could go into two different CGUs, one in virtue of the signal of understanding, and a new one to introduce its own content. Some felt that we should just include the acknowledgment function of the speech in the second CGU, while others advocated that this was not a problem since the role with respect to the first CGU was as an acknowledgment, and its role in the second is to add content. Perhaps marking these roles would solve the confusion. The second issue is that in a single turn, and hence in a single CGU, the speaker might make contributions that go into separate IUs. This presents a bigger problem to CGUs. We might need to add intentional analysis to CGUs so we can split things that would go into different intentional units.

Issue 2: Are CGU coding scores good?

κ scores for 4 dialogs differ a lot (see table 4, page 57). These κ scores showed a strong correlation with the amount of overlap in the dialogues. Verbmobil, which had none, due to the protocol that the participants followed, had the best κ scores, while Maptask, which had lots of overlap had the lowest κ scores. Overlapping speech is difficult to code for grounding units since

it is not as evident whether the overlapping speech is grounded and what it grounds.

The Toot dialogue, which was human-computer (all others were human-human) also received low κ scores. This could be due to the system talking for long stretches, including pauses to look up information on the web, and coders wanting to break it up more finely, or due to problems with coding misrecognitions of the system, or with the system not engaging in proper grounding behavior.

Another explanation for the divergence in κ scores could also be due to the number of coders who were familiar with the data. A number of those present had worked with the TRAINS corpus in the past, whereas few had seen the Toot dialogue.

We also found that not everyone listened to the dialogues to the same degree, which might account for the overall low κ scores. Some just listened once, while others played challenging stretches several times.

Although Maptask has the lowest κ scores, we felt that it is a good indicator of how far we must still go in clarifying the coding guidelines and grounding theory.

Issue 3: How independent should CGUs/IUs be from backwards/forwards

There is a section of the backward communicative function coding scheme which deals with *understanding acts*, which are clearly related to grounding. Moreover, several of the forward and backward communicative acts have to do with intentional/informational contents and structures. This is an issue that we briefly touched on, and will need to further address. Some of this was taken up in the plenary sessions.

Day 2: Report by Owen Rambow

What we did:

- We resolved 4 technical issues on CGU coding.
- We got feedback on the Japanese Map Task.
- We discussed a redefinition of the notion of "CGU".

The background for the day was the realization that people are having trouble understanding the theory behind CGUs, and thus are having trouble applying the coding guidelines.

Technical Issue 1: How do we tag self repairs?

The instructions in [Nakatani and Traum, 1998] stated that continuations and repairs should be put in the same CGU with their respective reparandums. However, a fresh-start need not be in the same CGU. A problematic middle-ground was "restart-continues", in which the repair starts at the beginning of the reparandum.

```
Example (Verbmobil):
      A.1.6  What day are good for you?
      A.1.7  What days are good for you?
```

Proposed solution: we will assume that the transcribed corpora we obtain are already annotated in such a way that speech repairs are marked, and will thus go in the same CGU.

Technical Issue 2: How do we treat cue words and fillers (e.g., *so*, *uhm*, *ah*) when they have been coded as own tokens?

The problem here is that it can be hard to tell how, if at all, these utterances fit in with grounding of surrounding material.

Proposed solution: the taggers do a 1-token lookahead and add to the most appropriate token (either the preceding one or the following one). The default is to choose the following token.

```
Example: Verbmobil A.1.1 should be in the same CGU as A.1.2
      A.1.1  So
      A.1.2  we should meet again
```

Technical Issue 3: How do we distinguish repair requests from clarifications and how does the distinction affect meso-level tagging?

The proposed solution is simply to clarify the definitions. A repair request occurs when H signals lack of understanding of S (either correct hearing or in terms of content). Repair requests are not tagged as embedded CGUs; rather, the CGU continues until understanding has been reached. A clarification request occurs when H did understand (both acoustically and in terms of content) S, but needs more (or different) information, for example in order to execute a domain task. Clarification requests are tagged as initiating a separate CGU (and perhaps acknowledging the prior one), as well as performing other content-level actions (such as a follow-up question).

Example: Verbmobil A.3.2-B.6.1

Technical Issue 4: How do we mark extra evidence of understanding of a CGU?

The proposed solution is that only enough evidence to consider something grounded will go in the prior CGU (modulo decisions about how to mark (different kinds of) acknowledgments, in general).

What Sort of CGUs do We Want?

The remaining difficulties in coding CGUs suggest that the definition of CGU, as presented in [Nakatani and Traum, 1998], may be too complex (and/or too vague) for use in coding. The main problem is to determine when information has entered the common ground (which is an important part of the

[Nakatani and Traum, 1998] definition). In the kind of corpora we have been analyzing mainly (exclusively), this issue is empirically observable through linguistic acknowledgments. In order to distinguish which elements of the [Nakatani and Traum, 1998], instructions and definitions are most useful, we looked at a large number of simple constructed examples. We found that there are, roughly, three types of acknowledgments:

1. Type 1 acknowledgments are explicit acknowledgments which contribute no new informational or intentional content (roughly speaking) of their own and only serve to acknowledge that the interlocutor has made a contribution and that the contribution has been understood. Put differently, they make no new dialog contribution other than the acknowledgment.

   In terms of micro-level analysis, Type 1 acknowledgments have only a backward-looking function, but no forward-looking function.

   ```
   Example: A: Why did you order a Martini?
            B: That is a very good question.
   ```

2. Type 2 acknowledgments implicitly acknowledge the interlocutor's contribution by performing a linguistic action which in normal discourse is expected as a response to (or at least plausibly follows) the previous contribution, but which itself also contributes informational or intentional content (roughly speaking).

   In terms of micro-level analysis, Type 2 acknowledgments have both a backward-looking function and a forward-looking function.

   ```
   Example: A: Why did you order a Martini?
            B: Because I like olives.
   ```

3. Type 3 acknowledgments are entirely implicit. For example, in an ongoing monologue there may be long stretches which never get explicitly acknowledged. The speaker will take the lack of protest (and assumptions about the quality of the communication channel) as sufficient evidence for acceptance by the interlocutor. An extreme case is, of course, written communication, in which case there can be no acknowledgment of any kind.

   In terms of micro-level analysis, Type 3 acknowledgments are manifested only by a forward-looking function (namely that of the next "discourse unit"); there is no backward-looking function that corresponds to them (since they have no explicit linguistic manifestation).

   ```
   Examples: A: Why did I order a Martini?
             A: Because I like olives.

             A: Why did you order a Martini?
             B: Where were you last night?
   ```

Discussion:

Type 1 acknowledgments should uncontroversially be included in the previous CGU (which they help to establish as common ground), since they have no independent function in the discourse. [Nakatani and Traum, 1998] suggested that Type 2 acknowledgments should also be included in the previous CGU to signal acceptance. Since they also contribute new material, they must also start a new CGU; they therefore invariably are contained in (at least) two CGUs. While the intersection of two CGUs is not in itself troubling, this definition was not deeply popular with coders, since this case in fact happens frequently in dialog, makes coding and reading the coding tedious, and can, so it was claimed by some, be easily retrieved automatically from previous micro-coding. It was therefore suggested that Type 2 acknowledgments should not be coded as belonging to the CGU they acknowledge (i.e., these cases of intersecting CGUs should be abandoned).

Finally, Type 3 acknowledgments proved the most controversial – there was no consensus whether they should be assumed in coding, or whether they even exist. From a practical (empirical) point of view, in the absence of acknowledgment of acceptance by the interlocutor, it is unclear what criteria to use to limit CGUs. From a theoretical point of view, it is unclear whether discourse participants consider unacknowledged material as part of the common ground or not. In response, it was pointed out that the criteria that should be used to delimit the CGUs in the case of no acknowledgments might be drawn from the requirements for the macro-analysis; this would give us roughly the notion of an atomic communicative intention (as has been developed in the literature). The issue of the theoretical motivation could be studied once annotated corpora are available (using these corpora), but should not be a bottleneck in devising a coding scheme.

During the post-lunch discussion in smaller break-out groups, one group discussed the relation between meso-analysis (CGUs) and micro-analysis in more detail. One hypothesis that was put forward was that, if we have a transcript properly annotated with micro-level codes, we could derive "stripped-down" CGUs semi-automatically, or even automatically: Type 1 acknowledgments, signaled by tokens with only backwards-looking function, would always be included in the current CGU. Type 2 acknowledgments, marked by the presence of a forward-looking function, would always trigger the closing of the current CGU and the opening of the next CGU. While this simple algorithm would not adequately account for repair requests and embedded subdialogs, it presumably could be extended in this sense. The motivation is the observation that entirely separate micro-, meso-, and macro-coding is undesirable because of the possibility of introducing sources of error and arbitrariness at each level of coding; inter-coder reliability would be higher if the coders were machines. Of course, such a "stripped-down" notion of CGU is only useful if it serves a purpose, for example the purpose of being the building block for the macro-analysis. Further study is required.

Day 2: Subgroup on CGU Coding of Japanese Maptask: Report by Yasuhiro Katagiri

This is a report on the preliminary attempt on the Common Grounding Unit (CGU) coding of the Japanese Maptask corpus in accordance with the Discourse Structure Coding Manual [Nakatani and Traum, 1998].

Focus

The purpose of this short exercise was to examine the feasibility of CGU coding with Japanese data, and see if there are Japanese dialogue phenomena that might be taken into consideration in further development of the CGU coding schemes. Seven Japanese speakers participated in the coding session, which took place during the Chiba DRI workshop. The possible issues we were aware of were:

- the complexities/difficulties in coding repairs/clarifications in terms of CGUs.
- Japanese acknowledgment expression *hai*, which can also be used as yes answers

Findings

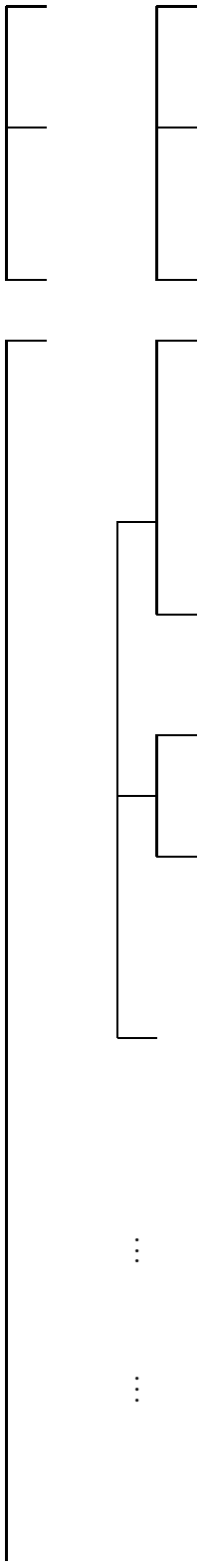
1. Grain size

   First, we found that there are two possible conceptions of CGUs which differ in grain sizes. We call them large CGUs and small CGUs.

   Large CGUs correspond roughly to transaction level coding. It was relatively easy to reach agreement among participants, so high coding reliability is expected at this level. But, information on the internal structures of interactions are lost at this level of coding.

   Small CGUs reflect finer details of grounding interactions between dialogue participants. But it was hard to reach agreement among participants. This was because (a) there are lots of points of overlapping speech in the data, (b) the dialogue contains many instances of repair requests and clarifications, which made it difficult to identify units of interaction.

   Example coding for each CGU conception is shown in Figure 16 together with English translations of the transcript. A line of text in the figure corresponds to an inter-pausal unit of speech (a stretch of speech bounded by silences of longer than 100msec). After G and F agreed upon their initial understanding of the problem situation, G started at G.3.2 to describe an initial portion of the route segment as "go down," which is subsequently refined interactively to "a little to the left and downward" and then "avoiding the campsite directly down from the start point." F.9.1 indicates F successfully identified (and followed) that path. This marks the end of a large CGU beginning at G.3.2. But, in the refinement process up to this, the interaction goes back and forth with repair requests and



L-CGU	S-CGU	Id	Time	Transcript
		G.1.1	00:01:904–00:02:468	dewa iidesuka Are you ready?
		F.1.1	00:02:508–00:02:704	hai Sure.
		G.2.1	00:03:356–00:04:644	ja mazu shuppatsuchitenni imasuyone You're at the starting point, right?
		F.2.1	00:04:708–00:04:864	hai Yes.
		G.3.1	00:05:368–00:05:900	etto(o) Um.
		G.3.2	00:06:240–00:06:644	ja mazu Ok, first,
		G.3.3	00:06:784–00:07:656	shitani mukatte go downward,
		G.3.4	00:07:764–00:08:684	shuppatsushitekuda * sai please start.
		F.3.1	00:08:540–00:10:268	* eeto(o) shitato iunoha So, what do you mean by downward?
		F.3.2	00:10:480–00:11:248	eeto mashitade Um, directly down.
		F.3.3	00:11:496–00:12:132	yoroshiidesuka Okay?
		G.4.1	00:12:324–00:13:808	achottohidariniitteshitani Uh, a little to the left and down.
		F.4.1	00:14:088–00:15:488	chotto hidari * niitte go to the left a bit.
		G.5.1	00:14:928–00:15:412	* itte go
		G.5.2	00:15:520–00:15:948	shi * tani down
		F.5.1	00:15:628–00:15:872	* shita down
		F.5.2	00:16:176–00:19:536	*1 eeto(o) shuppatsuchitenno sugu shitani ooto *2 kyanpujo Um, a campsite directly down from the starting point ...
		G.6.1	00:16:224–00:16:320	*1 i eh
	⋮	G.6.2	00:18:992–00:19:516	*2 tokyanpujou the campsite
		F.6.1	00:19:676–00:20:768	* ee sorewo sakeru(u) Um, avoid that
	⋮	G.7.1	00:19:768–00:19:984	* hai Yes
		G.7.2	00:20:864–00:21:616	sakeruyouni hai in such a way to avoid
		F.7.1	00:21:680–00:22:080	a hai Yes

L-CGU	S-CGU	Id	Time	Transcript
		G.8.1	00:22:528–00:23:592	hidarigawani sotte go aloing the left side, yes
	⋮	F.8.1	00:23:440–00:23:620	* hai Okay
		F.8.2	00:24:200–00:25:472	hai shitani(i) hai ikimashita(a) Okay, I went downward, ikay.
	⋮	G.9.1	00:25:788–00:26:384	de(e) Um,
		G.9.2	00:26:916–00:28:484	etto do dondon shitani ittekudasai go down further.
		F.9.1	00:28:432–00:28:624	hai Okay

Figure 16: Large and small CGUs

clarifications, and we have choices as to what constitute the units of refinement, or small CGUs. For example, we can take "left" and "down" as grounded either separately or together in the course of interaction. The judgment is made difficult partly because grounding sometimes occurs in overlapping speech.

2. Ambiguity in *hai*

   Figure 17 shows a large CGU coding of another fragment of the data. The fragment contains many instances of *hai*s, some of which mark the boundary between large CGUs, but many of them do not. Many probably even do not mark the boundary between small CGUs, as many of the *hai*s are best interpreted as simply showing acknowledgment that the hearer is attending to the speech. Accurate CGU coding might need to be performed with speech together with transcriptions, as speech prosody provides significant information toward disambiguation of these *hai*s.

Day 3: Report by Jennifer Chu-Carroll

Recap and Revision of CGU Analysis from Day 2

A revised proposal for CGU coding for questions that was proposed and investigated on day 2 (see Section 3.3.2) codes CGUs based on whether an utterance has a forward looking function, a backward looking function, or both. The question then is if we have a dialogue tagged with forward/backward looking functions, can we automatically derive the CGU codings for that dialogue? We decided to further investigate this problem by applying the new CGU coding rules developed for coding questions and answers to statements as well. This was discussed in a subgroup meeting in the afternoon and the preliminary results seem to show that although the concept could be extended, we need some more specific rules to group utterances that contribute to the same nucleus in

L-CGU	Id	Time	Transcript
	G.13.2	00:39:120-00:39:512	sorede then
	G.13.3	00:39:696-00:40:488	sokokara mata from there
	G.13.4	00:41:088-00:41:912	ee daitai Uh, about
	G.13.5	00:42:048-00:43:168	sansenchigurai ma three centi-meter ...
	G.13.6	00:43:496-00:45:008	tsumari ma furuisushagoyano(o) well, of the old water mill ...
	F.13.1	00:45:072-00:45:264	**hai** Uh-huh
	G.14.1	00:45:440-00:45:868	sokono its
	G.14.2	00:46:096-00:46:252	ma um
	G.14.3	00:46:388-00:47:648	eeto yo hidarigawano(o) uh, left side ...
	F.14.1	00:47:872-00:48:052	*__hai__ Uh-huh
	G.15.1	00:47:920-00:49:528	* shitagurain tokoromade tsuit-ara(a) somewhat down, when you reach around there,
	F.15.1	00:49:624-00:49:824	**hai** uh-huh,
	G.16.1	00:49:960-00:50:664	kondo sono then,
	G.16.2	00:51:336-00:52:496	furui suishagoyano(o) the old mill ...
	F.16.1	00:52:572-00:52:776	**hai** yes
	G.17.1	00:53:060-00:54:320	hidarigawani tsukuyouni(i) along its left side ...
	G.17.2	00:54:568-00:55:488	ueni agattekudasai go up
	F.17.1	00:55:396-00:56:500	eeto(o) suishagoya Um, the water mill
	F.17.2	00:56:640-00:57:532	no hidari desune left of it, right?
	G.18.1	00:57:540-00:57:732	**hai** yes
	G.18.2	00:57:896-00:58:128	**hai** Okay

Figure 17: Ambiguity in *hai*. (Occurrences of *hai* are marked by boldface letters.)

the same CGU Another criticism is that even if we were to devise a scheme so that we can automatically derive CGU codings from forward/backward looking functions, it seems ineffective to tag a complete set of forward/backward looking functions merely for the purpose of tagging CGUs. However, in practice, all we need to know for CGU coding is whether an utterance is forward or backward looking. In the case of backward looking utterances, we also need to know which utterance it refers to.

We applied this new CGU coding scheme to part of the Verbmobil dialogue (see section 3.2.2, a segment of which is shown below:

B.14.3	How about Great Scott
A.15.1	Sounds great except they have been out of business for a while
A.15.2	How about some other place
A.15.3	Let us just wander up Craig
A.15.4	and pick one we like that day
B.16.1	That sounds pretty good

Using the new CGU coding scheme, we would code the CGU units as follows: B.14.3 by itself, and A.15.1, A.15.2, A.15.3, A.15.4, and B.16.1 all in one unit. However, it was noted that in IU analysis, we may want to group B.14.3 and A.15.1 together to indicate rejection of a proposal and our CGU coding separates these two utterances into two units. Thus, a modification to the CGU coding scheme was proposed which would code utterances with both backward and forward looking functions by themselves in a CGU unit. The reason for this is that such utterances are likely to contribute at the intention level to both the CGU preceding it and the CGU after it. This revised coding scheme will then allow us to code the CGU units as follows: B.14.3 by itself, A.15.1 by itself, and the rest of the utterances in one unit. Based on this CGU coding, we can then combine B.14.3 and A.15.1 as one intentional unit without crossing CGU boundaries.

IU Analysis

A number of questions were raised during the discussion on IU analysis. First, in most cases coders disagree not with the intentional structure of the discourse, but with the level of granularity that one should code. For instance, some coders considered fixed CGUs #1-#6 to be in one flat segment *schedule date*, while others coded subsegments *ask about next week*, *ask about week after next*, etc. We agreed that determining the level of granularity for representing intentional structure is an unsolved problem, but brought up the possibility of revising the scoring system to take into account individual differences in the level of granularity coded, i.e., penalize situations in which two individuals coded the structure of the dialogue completely differently more than those in which two coding results differ only in terms of level of granularity. However, this revised scoring scheme was not pursued any further in the discussion.

The second issue that was discussed with respect to IU analysis is the representation of discourse structure. The current IU coding scheme specifies a tree structure for representing discourse intentions. From the representational

point of view, questions were raised as to whether or not the current tree structure is too rigid and therefore too limited in terms of its power in capturing discourse relations. A lattice structure was brought up as an alternative proposal, but again was not pursued. From the coding point of view, questions were raised as to whether or not the current tree structure is too complex for reliable coding agreement among coders. A proposal was brought up to merely code linear discourse segments without any embedding, but was rejected because we agreed that it is important to code hierarchical structure. In the end, we did not agree to make any changes to the current coding manual with respect to discourse structure.

We did reach one constructive agreement with respect to IU coding, regarding the coding of conventional openings and closings in conversations. Although the particular Verbmobil dialogue we examined had limited opening and closing, the other dialogues we analyzed in our coding exercises can be divided coarsely into *opening*, *make plan*, and *closing*. In these cases, some coders coded openings and closings at the top level in the discourse structure, while others coded them as embedded structures. We decided that for the purpose of intention analysis, it does not matter which way they are coded, as long as they are coded consistently among coders. As a result, we decided to incorporate the coding of openings and closings explicitly in the next draft of the coding manual.

Day 3 Subgroup on CGU Coding of TOOT Human-Computer Dialogues

Diane Litman, designer of the TOOT system led this subgroup which investigated the issues involved with using the coding scheme on the TOOT dialogue. There was some feeling from the coding exercise that this dialogue was very different from human-human task oriented dialogue and perhaps inappropriate to be used for this coding exercise (particularly for "debugging" the coding schemes. The main differences between TOOT dialogues (which involve a speech-dialogue interface to find Amtrak scheduling information via the WWW) are:

1. The dialogue system has a fairly limited range of allowable dialogue patterns, essentially equivalent to a FSA. Thus the dialogues and allowable intentional structures are forced to adhere to this structure regardless of the current desires of the human user.

2. The speech recognition/interpretation was not as good as between human users, and thus there were more rejections and misrecognitions.

3. Analysts were actually able to observe more about the grounding behavior than is usual in human-human dialogues, since there is also direct access to the way speech recognition system understood the words. Normally one can only tell how humans understand something indirectly, through feedback of how they react.

These differences led to a somewhat different dialogue and coding experience. However, a detailed examination of the dialogues and issues that came

up in coding them for CGUs indicated that there didn't seem to be any phenomena that were unique to human-computer interactions. The rejections and misrecognitions that occurred all had analogues in human-human dialogues, merely occurring with higher frequencies in the human-computer dialogues.

It was also suggested that this kind of CGU analysis, and particularly places where coders disagreed on the proper analysis might be helpful in assessing/debugging the design of the system's feedback mechanisms.

3.3.3 Synthesis

From the discussions on CGU coding, several goals emerged for having a level of CGU coding as part of DRI scheme:

1. Capture how content is added to the common ground

   One would like to be able to tell first, which bits of content get added to the common ground of dialogue participants, and secondarily how this happens, eg., at what point in the dialogue, and what acts or tokens play a role, and what that role is. The coding should also be as theory neutral as possible, so as to be able to test particular theories of grounding.

2. Be easy, reliable, and interesting to code

   Whatever coding scheme is devised, it must be possible for coders to learn to do it reliably, with a minimum of effort, in order to maximize the utility of coding efforts. Even if some aspects are "crucial" to actually getting grounding theory right, they may have to be left out of the coding scheme if they can't be easily coded. Likewise, if something can easily be done automatically, there's no need to get human coders to do it, so it doesn't need to be part of the coding scheme, but could be automatically derived for a more accurate account of grounding behavior.

3. Abstract away "messy" bits of dialogue (e.g., local repair, turn-taking, grounding) to something more like text, to build as the basis of IU structure.

   Lots of Intentional/informational discourse structure theories rely on a single point of view of the structure, which is either the point of view of the writer, for monologue, or the assumed common point of view of dialogue participants. A problem for these theories is that at a low level, it is very clear that dialogue participants have very different points of view about what is happening. The hope is that these differences (at least about what is being said and meant) will occur within CGUs, and by the time of achieving complete CGUs, a single point of view could be assumed, and one could then look at how this established content could be combined into a single intentional/informational structure.

4. Be consistent with Forward/Backward and IU coding

   There are several types of consistency. First, since Forward/Backward currently contains coding dimensions for some aspects of grounding —

i.e., the "understanding acts", it would be good to avoid duplicating the effort, and more importantly not coming up with something inconsistent. Also, there's the issue of whether the CGU structures derived can either make use of or be useful for forward/backward and IU coding. So this category can be split up into sub-principles as follows:

   (a) combine with understanding level acts into one coherent view of grounding

   (b) possibly use forward/backward principles in constructing CGUs

   (c) possibly make CGUs useful units for forward backward act principles to consider

   (d) possibly make use of intentional/informational principles in constructing CGUs

   (e) (same as 3, above, though possibly broader and less committal) possibly make CGUs useful units for intentional/informational principles to consider

5. make clear what intentional/informational content is added to common ground by the CGU.

   This is a derived desire, coming from both (1) and (3,4) - if one is to actually use CGUs for something, one needs to know what content was added to common ground, not just that *some* content was.

Not everyone in the group was concerned with all of these goals. Some expressed a concern only with grounding, and are not at all with intentional structure. Others didn't really care about grounding per se, but only how it could be used as a tool in simplifying other levels of analysis, such as intentional structure. It did not seem that anyone actually objected to any of these five concerns, however, although specific proposals might lead to furtherance of one of the objectives at the expense of another. Such proposals might thus be objectionable to those who do care about the neglected goals. Hopefully we can devise a scheme which adequately satisfies all of the objectives. If this is not possible, we might have to have two, orthogonal "meso-level" notions, one common ground coding scheme, and one "IU starting point scheme" which will be somewhat orthogonal.

Proposals:

There are several main ways to address goal (5), so that the observable CGU grounded content is satisfactory for goals (1) and/or (3,4).

a. change the policy about which tokens are added to the unit, so that the simple heuristic of examining the content of the included tokens will more accurately reflect the actual content. Likewise for calculating bracketing endpoints of CGUs for use in assigning forward/backward act and intentional structures.

b. specify the grounding function (e.g., adding content, repairing content, acknowledging, cancelling, etc) of each token with respect to the unit, instead of just listing the token itself. This will allow a better heuristic to be used to judge the content, and perhaps segmentation boundaries. Just having the grounding functions themselves will help, but they may also allow an automated summarization routine to compute a text-like content based on the tagged data, which would be a better input for intentional structure (and perhaps forward/backward acts).

c. use a different kind of basic content for CGUs: forward/backward acts rather than primitive tokens. This would also help determine what the content was, since only some of the set of acts associated with a token would be added to the CGU when the token plays a grounding relation. This also allows for a more flexible and abstract notion of intentional structure, but requires forward/backward coding to be done first.

d. specify through other means the understood grounded content. This was the approach taken, to a limited degree in [Nakatani and Traum, 1998], by inviting subjects to mark down the intentional content result of CGUs, following ideas developed in [Nakatani *et al.*, 1995]. The advantage is that it allows coders to specify in as much detail as they choose what content is added to common ground as the result of a unit. The disadvantage is that it may be difficult to compare these contents or automatically analyze them.

One discussed proposal amounts to a version of (b), in which one separates the acknowledgment function of some tokens, from the presentation (or content-adding) function of others. This will help with goal (5) and thereby (1), (3) and (4), though perhaps at the expense of (2), just in virtue of trying to code more things. This can probably be fairly easily overcome with some simple coding tools or even policies.

Another discussed proposal, seems to have several different (and perhaps conflicting?) motivations. The proposal itself is along the lines of (a), above, specifically in terms of not including answers to questions in the CGU of the question.

Some motivations to be related to goal (2): since answers would be a backward relation provided by FB group coding, and since an answer would generally ground the question (unless it was already grounded), there was no need to specifically code the acknowledgment relation (or put the answer token in the CGU of the question). The grounding relation itself, crucially a part of goal (1), would not be compromised, since the acknowledgment function could be derived, when necessary, from the answer relation. An additional benefit for goal (2), which was perhaps not explicitly mentioned in the discussion of this proposal, is that it would eliminate the need to decide how much of multi-token answers should be put in the CGU of the question, which would tend to increase coder reliability (on the other hand, for some goals of (1), people might actually be interested in that very question, so in that case this coding

wouldn't help them, though they could always add on a finer layer of coding to address this).

Another motivation was according to goal (1): a feeling that questions are grounded differently from other forward acts such as statements, and therefore should be coded differently.

Yet a third motivation seemed to be with respect to (5), particularly as applied to (4): The idea that since the answer had different forward/backward functions, and different intentional structure, and one might want to put a boundary of some sort between question and answer, it was important to have them in separate units.

Some have also expressed the feeling that these two proposals are notational variants of each other: if one has the answer relations and perhaps other forward/backward codings (perhaps only at a very abstract level), one could use the latter proposal, whereas if one did not have this level of coding, one could use something like the former proposal and mark acknowledgments.

3.4 Summary

This meeting was very much a preliminary rather than decisive meeting on discourse structure in dialogue. Several further aspects must be addressed before coming to agreement on a general-purpose coding scheme for discourse structure in dialogue. While the reliability results presented in Section 3.2 are already close to acceptable, and might achieve still higher results with additional training, there was a general consensus that further modifications along the lines discussed above, would produce better coding schemes. In particular, the relationship between these codings and related parts of the Forward/Backward coding schemes should be explored, and more codings of actual dialogues be performed to assess the concrete results of the above proposal modifications.

4 List of Participants of the Open Session at Chiba University

Masahiro Araki	University of Kyoto
Timothy J. Baldwin	Tokyo Institute of Technology
Ellen Gurman Bard	University of Edinburgh
Jean Carletta	University of Edinburgh
Key-Sun Choi	KAIST
Jennifer Chu-Carroll	Bell Labs.
Mark Core	University of Rochester
Morena Danieli	CSELT
Yasuharu Den	Nara Advanced Institute of Science and Technology
Barbara Di Eugenio	Univ of Illinois at Chicago
Mika Enomoto	Chiba University
Satoru Fujimura	Sony D21 Labs.
Tsutomu Fujinami	Japan Advanced Institute of Science and Technology
Sanae Fujita	Nara Advanced Institute of Science and Technology
Yoshiaki Fukada	Waseda University
Jun-ichi Fukumoto	Oki Electric Industry CO., Ltd
Yuko Goto	Stanford University
Pat Healey	ATR Media Integration and Communications Research Labs.
Peter Heeman	Oregon Graduate Institue of Science and Technology
Julia Hirschberg	AT&T Labs.
Akira Ichikawa	Chiba University
Masato Ishizaki	Japan Advanced Institute of Science and Technology
Kristiina Jokinen	ATR Interpreting Telecommunications Research Labs.
Hideki Kashioka	ATR Interpreting Telecommunications Research Labs.
Masahito Kawamori	NTT Basic Research Labs.
Hideaki Kikuchi	Waseda University
Naomi Hamazaki	Chukyo University
Koiti Hasida	ElectroTechnical Lab.
Kiyota Hashimoto	Seiwa University
Shoji Hayakawa	Fujitsu Co.
Yasuo Horiuchi	Chiba University
Hitoshi Iida	ATR Interpreting Telecommunications Research Labs.
Riyoko Ikeda	National Language Research Institute
Toshihiko Itoh	Toyohashi Univeristy of Technology
Susanne J. Jekat	University of Hamburg
Dan Jurafsky	University of Colorado
Yuko Kasahara	University of Electro-Communications
Hiroshi Kanayama	University of Tokyo
Takuya Kaneko	Keio University
Hideko Kashiwazaki	Tokyo Institute of Technology
Kazuto Kasuya	Omron Co.

Yasuhiro Katagiri	ATR Media Integration and Communications Research Labs.
Takeshi Kawabata	NTT Basic Research Labs.
Satoshi Kobayashi	Shizuoka University
Hisashi Komatsu	Hiroshima City University
Hanae Koiso	National Language Research Institute
Yuichi Kojima	Ricoh Co., Ltd
Tomoko Kumagai	National Language Research Institute
Sadao Kurohashi	Kyoto University
Teishu Lee	Waseda University
Lori Levin	CMU
Diane Litman	AT&T Labs.
Kikuo Maekawa	National Language Research Institute
Nobuo Masaki	ATR Human Information Processing Research Labs.
Noboru Miyazaki	NTT Basic Research Labs.
Johanna Moore	University of Edinburgh
Makiko Naka	Chiba University
Kenjiro Nagano	Nagaoka University of Technology
Tomoko Nakamizo	Tokyo University of Foreign Studies
Yukiko Nakano	NTT Communications and Information Labs.
Christine H. Nakatani	Bell Labs.
Shu Nakazato	Meio University
Hiroaki Noguchi	Nara Advanced Institute of Science and Technology
Tadashi Nomoto	Hitachi Basic Research Labs.
Takashi Ninomiya	University of Tokyo
Hiroyuki Nishizawa	Tokiwa University
David G. Novick	Eurisco
Shigeki Ohira	Waseda University
Yoshimitsu Ozaki	National Language Research Institute
Owen Rambow	CoGenTex
Norbert Reithinger	DFKI
Tomoko Sasaki	National Language Research Labs.
Akira Saso	Tokyo Denki University
Nozusawa Shiho	Keio University
Atsushi Shimojima	Japan Advanced Institute of Science and Technology
Tsubasa Shinozaki	NTT Human Interface Labs.
Ikuko Shiotsubo	Kochi University
Teresa Sikorski	University of Rochester
Hidetoshi Shirai	Chukyo University
Michael Strube	University of Pennsylvania
Yoshiaki Sugaya	ATR Interpreting Telecommunications Research Labs.
Yosuke Sugita	Waseda University
Hiroyuki Suzuki	Panasonic
Kazuhiko Tajima	Sony D21 Labs.

Kazuhiro Takeuchi	Nara Advanced Institute of Science and Technology
Masafumi Tamoto	NTT Basic Research Labs.
Hideki Tanaka	ATR Interpreting Telecommunications Research Labs.
Yuka Tateishi	University of Tokyo
Wataru Tsukahara	University of Tokyo
Syun Tutiya	Chiba University
David Traum	University of Maryland
Masao Uchiyama	Shinshu University
Jennifer J. Venditti	Ohio State University
Yoichi Yamashita	Ritsumeikan University
Hiroyuki Yano	Communition Research Lab.
Tomohiko Yoshida	Chiba University
Koichi Yoshimura	Kanazawa University
Takashi Yoshimura	ElectroTechnical Lab.
Gregory Ward	Northwestern University
Michiko Watanabe	University of Tokyo
Yasuhiro Watanabe	Omron Co.
Keiko Watanuki	Sharp Co.

References

[Alexandersson *et al.*, 1997] Jan Alexandersson, Bianka Buschbeck-Wolf, Tsutomu Fujinami, Elisabeth Maier, Norbert Reithinger, Birte Schmitz, and Melanie Siegel. Dialogue acts in verbmobil-2. Technical Report 204, DFKI GmbH Saarbrücken and Universität Stuttgart and Technische Universität Berlin and Universität des Saarlandes, 1997. Also available at `http://www.dfki.de/cgi-bin/verbmobil/htbin/-doc-access.cgi`.

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

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

[Carletta *et al.*, 1997a] Jean Carletta, Nils Dahlbäck, Norbert Reithinger, and Marilyn A. Walker, editors. *Standards for Dialogue Coding in Natural Language Processing*, Schloß Dagstuhl, 1997. Seminar Report 167, also available at `http://www.dfki.de/dri/`.

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

[Clark and Marshall, 1981] Herbert H. Clark and Catherine R. Marshall. Definite reference and mutual knowledge. In Aravind K. Joshi, Bonnie L. Webber, and Ivan A. Sag, editors, *Elements of Discourse Understanding*. Cambridge University Press, 1981. Also appears as Chapter 1 in [Clark, 1992].

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

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

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

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

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

[Godfrey *et al.*, 1992] J. Godfrey, E. Holliman, and J.McDaniel. SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings of ICASSP*, pages 517–520, San Francisco, 1992. IEEE.

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

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

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

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

[Heritage and Roth, 1995] John C. Heritage and Andrew L. Roth. Grammar and institution: Questions and questioning in the broadcast news interview. *Research on Language and Social Interaction*, 28(1):1–60, 1995.

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

[Jekat *et al.*, 1997] Susanne Jekat, Heike Tappe, Heiko Gerlach, and Thomas Schöllhammer. Dialogue interpreting: Data and analysis. Verbmobil-Report 65, Universität Hamburg, 1997.

[Jurafsky *et al.*, 1997] Daniel Jurafsky, Elizabeth Shriberg, and Debra Biasca. Switchboard: SWBD-DAMSL labeling project coder's manual, draft 13. Technical Report 97-02, University of Colorado Institute for Cognitive Studies, 1997. Also available at `http://stripe.colorado.edu/~jurafsky/manual.august1.html`.

[Kowtko *et al.*, 1991] Jacqueline C. Kowtko, S. Isard, and G. Doherty. Conversational games within dialogue. In *Proceedings of the ESPRIT Workshop on Discourse Coherence*, 1991.

[Labov and Fanshel, 1977] William Labov and David Fanshel. *Therapeutic Discourse*. Academic Press, New York, 1977.

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

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

[Nakatani and Traum, 1998] Christine H. Nakatani and David R. Traum. Discourse structure coding manual. Unpublished draft manuscript, 1998.

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

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

[Pierrehumbert, 1980] J. B. Pierrehumbert. The phonology and phonetics of english intonation. Doctoral dissertation, Massachusetts Institute of Technology, 1980.

[Pollack, 1986] Martha E. Pollack. *Inferring Domain Plans in Question-Answering*. PhD thesis, University of Pennsylvania, 1986.

[Quirk *et al.*, 1985] Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. *A Comprehensive Grammar of the English Language*. Longman, London, 1985.

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

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

[Traum, 1998] David R. Traum. Notes on dialogue structure. Unpublished manuscript, 1998.

[Weber, 1993] Elizabeth G. Weber. *Varieties of Questions in English Conversation*. John Benjamins, Amsterdam, 1993.