Report on the Dagstuhl-Seminar

Standards for Dialogue Coding in Natural Language Processing

Organizers:

Jean Carletta (University of Edinburgh)
Nils Dahlbäck (Linköping University)
Norbert Reithinger (DFKI)
Marilyn A. Walker (AT&T)

February, 3.–7., 1997

This report can be retrieved from the seminar homepage `http://www.dfki.de/dri/` or from the Dagstuhl homepage `http://www.dag.uni-sb.de/ENG/`, where you can also find information about the *Internationales Begegnungs- und Forschungszentrum für Informatik SchloßDagstuhl*

Contents

1 Introduction

Norbert Reithinger, DFKI

During the last years, corpus based approaches have gained significant importance in the field of natural language processing. Currently, large corpora for many different languages are being collected all over the world. In order for these data to be useful for training and for testing implemented systems, the corpora must be annotated in various ways. The Discourse Resource Initiative (DRI) is an effort to assemble discourse resources in support of discourse research and applications.[1] This seminar was the second in a series of DRI workshops with the goal to develop a standard to annotate corpora for semantic/pragmatic and discourse features.

As for the previous workshop in Philadelphia, PA., the participating 43 researchers had to do some homework, namely annotating different texts with the schema defined in the first workshop, before coming to the workshop.

During the workshop, the audience split in five groups to discuss different aspects of the annotations, namely coreference, forward looking functions, backward looking functions, segmentation, and information level and information status. At the end of the workshop the results were presented and discussed. This report contains the summaries from these five groups. During the workshop, additional groups met, e.g. to discuss tools for annotation and to discuss interactions between forward/backward looking functions and coreference.

There were also five plenary talks by

- James Allen: Towards a Standard for Annotating the Structure of Dialogue
- Hans Dybkjær: Dialogue Annotation in Europe
- Masato Ishizaki & Syun Tutiya: An Attempt to Standardize Discourse Tags in Japanese
- Lynette Hirschman & Rebecca Passoneau: Coreference Annotation
- Marc Vilain: A Model-Theoretic Coreference Scoring Scheme

that gave an overview of the last workshop, presented ongoing work in Europe and Japan, and addressed aspects of coreference annotation and scoring.

[1] See also `http://www.georgetown.edu/luperfoy/Discourse-Treebank/dri-home.html`

2 Summary of the Coreference Group

Rebecca Passoneau, Columbia University

2.1 Group Membership

Participants who attended the Dagstuhl meeting, plus Lynette Hirschman, co-organizer with Rebecca Passonneau, are listed below. Lynette Hirschman was unable to attend due to an injury, but prepared much of the material that was presented or discussed at the meeting.

Lars Ahrenberg	lah@ida.liu.se
Dan Cristea	dcristea@infoiasi.ro
Dick Crouch	crouch@signal.dra.hmg.gb
Lynette Hirschman	lynette@mitre.org
David Milward	milward@cam.sri.com
Rebecca Passonneau	becky@cs.columbia.edu
Laurent Romary	romary@loria.fr
Michael Strube	strube@coling.uni-freiburg.de
Marc Vilain	mbv@.mitre.org
Bonnie Webber	bonnie@central.cis.upenn.edu

2.2 Homework Results

The group annotated 5 short discourses. Two were annotated for coreference using the MUC Coreference Specification (MUCCS), prepared by Lynette Hirschman. These were a short (N=176 word tokens) Wall Street Journal article and a ABC news wire item (N=403). Two dialogues were annotated for coreference using the DRAMA (Discourse Reference Annotation for Multiple Applications) annotation, documented by Rebecca Passonneau. The same dialogues, a Trains dialogue (N=221 words; collected at U. Rochester) and a Coconut dialogue (N=311; collected at U. Pitt) were also coded by the dialogue act coding subgroup. Finally, the fifth discourse, a spoken narrative (N=411; from Wallace Chafe's "Pear Stories" [1980]), was coded for other relations among referents besides identity of reference, primarily exemplifying what is referred to as meronymy in WordNet [Miller et al., 1990] (e.g., set/subset, set/member, part/whole).

The several goals were to compare reliability of coreference coding across different types of discourse, to compare two annotation specifications for coreference coding, and to begin to examine the annotation, and eventually scoring methods, for other referential relations besides identity of reference.

The method for scoring identity of reference is the one presented in Vilain et al. [1995]. Briefly, identity of reference is a transitive, symmetric, reflexive relation, hence defines an equivalence class. Two codings of identify of reference for the same discourse are compared by comparing their equivalence classes in a fashion analogous to more familiar applications of precision and recall. An equivalence class is treated as a spanning tree whose nodes are the class members. Given an equivalence class in the answer key that is a superset of the corresponding equivalence class(es) in the response, the response is missing some of the equivalence relation arcs, and therefore recall is less than perfect. Conversely, given an equivalence class in the answer key that is a subset of the corresonding equivalence class in the response, the response contains too many equivalence relation arcs and precision is therefore less than perfect. For the actual computation of recall and precision, see Vilain et al. [1995].

All results on the coreference codings are scored using the method mentioned above. The results on the two MUCCS discourses were quite comparable to one another. There were five responses scored for the current workshop, compared with eight from the first DRI workshop held at U. Penn. With the qualification that there may be two few data points to provide a fair comparison, there seems to have been some improvement in scores using MUCCS. In the results on newswire text presented at the Penn workshop, there was a greater spread in both recall (between roughly .3 to .95, but mostly clustered from .6 to .8) and precision (between .75 and .95). At Dagstuhl, the recall was ranged from .54 and .85, but mostly clustered around .8; precision ranged from .80 and .95. This may simply reflect differences in the Penn and Dagstuhl texts, or may reflect improvements in the MUCCS annotation specification.

For both Dagstuhl dialogues, the results showed two quite distinct clusters of results. That is, 4 of the 6 scores on the Trains dialogue were clustered together with recall and precision both at or above .80; two of the scores had similarly high precision, but relatively low recall at around .55. For the Coconut dialogue, 3 of the 5 scores were clustered at around .7 recall and precision while 2 where clustered at around a low of .2 recall and about .6 precision. The low recall clusters were from the same two coders on both dialogues. Lower scores were due primarily to the omission of annotations for first and second person pronouns, and demonstrative pronouns.

If only the good Dagstuhl clusters are considered, then the results on the Dagstuhl Trains dialogue were comparable to the Penn results, with higher scores and slightly less spread on Trains than on newswire. Scores were less good on the Coconut dialogues. Three problems in particular seemed to recur across coders in the Coconut annotations. First, both of the dialogue participants had a number of chairs that were considered in solving the room-furnishing task. References to the various sets and subsets of chairs were consistently confused across coders. Second, despite the ease of interpreting personal pronouns, annotators often made errors in coding "I" and "you" across turns. Failure to notice a turn shift might account for both problems, but it is not clear why tracking turn-taking would be more difficult in Coconut than in Trains dialogues. Finally, there were numerous errors in coding antecedents of the demonstrative pronoun "that." This probably reflects an inherent difficulty in resolving discourse deixis.

Regarding the lower scores on Coconut, it should be noted that coding of dialogue acts was also less consistent on Coconut than on Trains. In the dialogue act coding group it was conjectured that the machine-mediated Coconut dialogues are somewhat less natural because the person taking the current turn sees the entire preceding turn while formulating a response.

Audio files were available for the Trains dialogue and were used by some of the annotators. There was no discernible pattern of scores associated with listening to or not listening to the audio.

The Pear narrative answer key contained 5 types of bridging inferences out of 7 possible ones documented in the DRAMA manual. These included one set relation: subset. However, no attempt was made to use the transitive closure of the subset relation to score results. Recall and precision for all bridging inferences was scored as follows. A particular type of bridging inference B was represented as a relation between a particular word or phrase token t in the discourse and an equivalence class e of referential identity (<t, B, e>). Take for example the utterance "there are only two", where the token "two" (token id=x) refers to two baskets that are a subset of a set of three baskets (equivalence class e.1) that have been mentioned several times earlier. To represent this particular subset relation, the answer key will contain the 3-tuple <x, subset, e.1>. The output of the tool used for annotating bridging inferences was mapped to this type of representation. For each of the 7 types of possible bridging inferences, separate recall and precision scores were computed.

In general, precision was better than recall for the bridging inferences. In DRAMA, possessive pronouns and genitive NPs are annotated as providing a bridge from the referent of the NP (e.g., from the bicycle referent in "bicycle of the boy's") to the referent of the possessive pronoun or genitive NP (e.g., to the boy to whom the bicycle is inferred to belong). Precision for this relation was perfect, recall was only .47. Recall could easily be improved by pre-tagging all possessive pronouns and presumed genitive NPs. The two other types of inferences that occurred at all were 'subset' and 'member'. Recall was somewhat better for the 'member' relation at .57 compared with .48 for the 'subset' relation; precision was quite a bit better for the 'subset' relation at .84 as compared with .65 for the 'member' relation. Here recall could also potentially be improved, e.g., by using WordNet to pre-tag possible instances of these relations; WordNet encodes meronymy between noun concepts, which is defined to include various kinds of constituency such as member/set, set/superset, part/whole (and their inverses). The answer key contained very few bridging inferences involving propositional antecedents, or implicit arguments and implicit partitives, and scores for both were low. No causal relations or part/whole relations were in the answer key, and none were in the responses.

In sum, precision scores for the bridging inferences are in general comparable to precision for identify of reference. Recall is poorer, but could potentially be improved by automated pre-processing. The coder's task would then involve accepting or rejecting the pre-tagged items, and looking for items that the pre-processor might have missed. Even for annotating implicit arguments, there is a potential for automated support of the annotation task. For example, the noun entries in the COMLEX [Grishman et al., 1994] syntactic resource contain some subcategorization information which could help identify potential cases of implicit arguments. There was little discussion of scoring bridging inferences, but Marc Vilain presented a discussion of problems in scoring transitive relations such as the subset relation. There was a short working session on this topic that included some members of the coreference and dialogue acts groups.

The primary role of the homework was to familiarize members of a working group with a common data set, and common annotation specifications. While the homework results cannot be considered to provide reliable measures of differences in language varieties, of improvements in coding specifications, and so on, they do indicate some interesting trends, and potential generalizations. Thus, the higher scores on annotating coreference in Trains dialogues (and in a direction-giving task from the Penn workshop) than on newswire may reflect differences in the language use: more concrete referents, more specific, immediate locations, and fewer referents that change over time in the face-to-face problem-solving dialogue versus more abstractions, more complex events, and semantically more complex language in newswire. Also, refinements in the MUC Coreference Specification may have led to improvements in the scores on newswire.

2.3 Resolved coding issues

Among the co-reference group participants there was implicit agreement that identity of reference for relevant noun phrases should be annotated, and that the evaluation metric presented in Vilain et al. [1995] was a useful way to quantify results comparing new codings against an a priori target or answer key. There was significant disagreement on only one linguistic phenomenon regarding whether it fell within the category of referential identity, as described below at (A). In addition, there was general agreement (B) that referential properties of constituents other than noun phrases should be annotated, (C) that the annotation should distinguish between identity of reference and discourse anaphoric relations, and (D) that other referential relations should be annotated, including type/instance relations as well as the types of bridging inferences documented in the DRAMA manual. These issues are taken up in turn in the remainder of this section.

The general terminology adopted here is that the linguistic expressions to be annotated for identity of reference and other features are referred to as markables. The various relations among the referents of markables, such as identify of reference or subset, are referred to as links. A link relation has a target (i.e., from the domain) and a source (in the range). As noted below, there is potentially some indeterminacy as to whether the source and target are linguistic tokens (e.g., in the discourse anaphoric link type) or their denotations (e.g., in the identify of reference link type). This may be predictable on the basis of the semantics of the link type.

(A) The main disagreement concerned a phenomenon associated with copula sentences, i.e., where the propositional content of the sentence is an assertion of equality between the denotation of the sentential subject and that of a noun phrase complement. The question raised was essentially whether the propositional link of equality between the subject referent and the complement referent should be equated with (and annotated in the same as) links of inferred referential identity. The following example is adapted from MUCCS V4.0, Oct 19, 1996.

(1) "Henry Higgins is the president of Dreamy Detergents"

This phenomenon is handled differently in MUCSS and DRAMA: for MUCSS, an asserted equality is treated as equivalent to referential identity while in DRAMA the subject and complement NPs are kept referentially distinct. Several approaches were discussed within the working group. Some of the problems raised, e.g., in preserving the semantics of the equivalence relation of referential identity, are discussed in the MUCSS documentation, section 3.2 "Terminology for Mark-Up". In some cases, e.g., the above example, there is also an interaction with the semantics of relational nouns like "president." It was decided that the disagreement was not resolvable at the current meeting, that pragmatic annotation solutions exist (e.g., propositional equality could in practice be segregated within any annotation convention, thus allowing any future analyst of the same data to treat these cases as desired), and further discussion was tabled in order to progress on other fronts.

(B) The sets of markables in the two annotation schemes largely overlap, but focus primarily on various types of noun phrases and what are referred to in DRAMA as noun phrase surrogates. (Noun phrase surrogates are constituents whose internal structure is not that of a prototypical noun phrase, such as what are sometimes referred to as headless NPs, but which function syntactically like NPs.) It was noted that the two annotation schemes differ as to the extent of the markable that must be annotated. For practical reasons, the MUCCS scheme allows a markable to consist in the noun group spanning the determiner up to the head, but potentially omitting post-modifiers. The DRAMA scheme specifies the markable NP to include all restrictive modifiers. It was agreed that regardless of the syntactic type of markable, a community-wide annotation specification should indicate the extent of the markable—possibly with options—as well as its type. In particular, neither annotation scheme addressed the problem of discontinuous markables. The classic examples from French involve clitics, e.g., the clitic "en" and the numeral/determiner "trois" in the following example should constitute a single markable:

(2) Elles en avaient trois.
Gloss: they-(feminine gender) of-it/them had-plural three
Trans: They had three of them.

The group identified four general classes of markables, and noted that there may be some variation across languages, and certainly differences of detail, in a complete annotation specification. The four classes to be addressed are i) noun phrases, ii) clauses and/or tensed verb phrases, iii) constituents that are syntactically intermediate between i) and ii) (e.g., gerundive and infinitival phrases), and finally, iv) super-sentential units.

Independently of any links among markables, each markable is to be annotated with three attributes. The "key" attribute serves as a token identifier. The "type" attribute is to have a value from the set individual, set, kind, property, event, generic. Assigning the value of such a "type" attribute can, for example, serve the function of distinguishing between the use of a phrase like "the president" to denote an individual (e.g., Bill Clinton) versus a type (e.g., the elective office). Whether the type attribute can be defined precisely and assigned reliably remains to be worked out, and will be addressed in a post-workshop cycle of annotation agreed to at the Dagstuhl meeting (cf. section 2.5).

(C) Consider the three sample texts below illustrating three cases of identify of reference, but by different linguistic means.

(3) a. Joachim Vandenburg bought a castle from Frederick Rothschild.
b. [Joachim Vandenburg] was pleased with his new acquisition.

(4) a. Joachim Vandenburg bought a castle from Frederick Rothschild.
b. [He] was pleased with his new acquisition.

(5) (Joachim Vandenburg enters the room. Person A addresses person B, nodding towards Joachim, and says)
[He] just bought a new castle.

All three examples illustrate markables (the bracketed expressions) that are used to refer to a person named Joachim Vandenburg. In (3) the relevant markable is a proper name whose interpretation does not depend on the prior linguistic context. The referential identify of the two uses of the proper name "Joachim Vandenburg" in (3a) and (3b) need to be annotated, but there should be no anaphoric link between them. In contrast, the markable in (4) is an expression that in this context is discourse anaphoric: the interpretation of the pronoun "he" depends on the contextual availability of a previously mentioned referent or referents, and in this case, there are two available referents, one of which is the antecedent or "source". Finally, example (5) illustrates a deictic use of the pronoun "he", where the "source" is in the situational context, rather than being linguistically evoked.

(D) A hierarchy of link types was agreed upon, and is represented in Figure 1. Attributes of links would include the link type, with values from Figure 1; the target, whose value would be the "key" (token identifier) of the target markable; and the source, whose value would be the "key" of the source markable, or a list of such keys, e.g., where a plural pronoun is linked to several antecedent

```
Referential links
  |
  +--- Referential identity
  |
  +--- Bridging links
         |
         +---Component relations (part/whole, etc.)
         |
         +---Lexical role-filling
         |
         +---Set relations
                 |
                 +--- subset/superset
                 |
                 +--- membership/containment
                 |
                 +--- co-membership
```

Figure 1: Hierarchy of Link Types

individuals.

During the first working session, it was proposed that the target type and the source type should also be annotated. Values would range over the same values specified above, in section (B) for the type of a markable: individual, set, property, and so on. This would help distinguish between cases such as the two types of referential relation illustrated in (6) (Note: This type of example was discussed by Nunberg in his examination of the semantics of lexical items such as president and newspaper, where the referent could be the individual or office in the case of a use of 'president', or, in the case of a use of 'newspaper, the institution or its product. Nunberg used the possibility of anaphoric reference to the different guises of a concept as a criterion for distinguishing predictable type shifting from true polysemy.)

(6) a. [1: The president] is sworn in 2 months after an election.
b. [2: He] does not officially take office until then.
c. But when [3: he] does, he and his VP Al Gore plan to celebrate.

The relevant markables in (6) have been bracketed and assigned the keys 1: through 3:. (Note that the notation shown here does not reflect the actual annotation language, which is to be implemented in SGML.) The markables 1: and 2: would be linked by identity of reference. Then according to the first proposal, the target type and source type for the link from 2: to 1: would presumably both be "kind." In contrast, for the link between 3: and 2:, the link type would again be identity, the source type would "kind", but the target type would be "individual." However, in a subsequent working session it was resolved that the target or source type would not be a property of the link, but a property of the markable. By this proposal, the two links between 2: and 1:, and between 3: and 2:, would both be of the type identity. The semantic difference would be captured by the type attributes of the three markables: the markables 1: and 2: would have the value "kind" for their type attribute whereas the markable 3: would have the value "individual."

Finally, in addition to link type, target and source, links would have a fourth "relation". The three possible values are external, e.g., for deictic reference, as in (5); anaphoric, e.g., for the two links as in (4) and (6); and internal. Internal would apply to cases where the referent of a markable is resolved through some reasoning process involving the prior linguistic context, but where the relation is not anaphoric. In (7), for example, it can be inferred that "Matthew" is linked by identity of reference to "an old friend", but the relation is neither anaphoric nor external, it is internal to the linguistic context.

(7) a. Susan went to spend the day with an old friend.
b. She enjoyed seeing Matthew again after all these years.

In summary, it was agreed that the combinatoric possibilities of the various annotation elements, their attributes and attribute values had not yet been fully elaborated, and that therefore the utility and convenience of the proposal remained to be evaluated. Moving towards such an evaluation would be one goal of the post-workshop annotation homework mentioned above in section 2.3.(C), and outlined below in section VI.

2.4 SGML Representation

Several proposals for encoding the annotation in an SGML annotation language were examined. The markup language for MUCCS uses "COREF" as the element tag for markables, assigns a type attribute the value "IDENT" for identity of reference, and uses a REF attribute to point from a markable to its antecedent:

```
<COREF ID="100">IBM</COREF> announced a dividend.
<COREF ID="101" "TYPE="IDENT" REF="100">It</COREF> showed a hefty
 fourth quarter profit.
```

The markup language for DRAMA, which was discussed in presentations at the workshop but which is not documented in the manual, assigns element tags to all markables, whether a subsequent expression is used to corefer with it or not. All marked elements get an ID attribute and a REFINDEX attribute. The REFINDEX attribute is assigned a distinct value from any previous values if the referent of the markable is new; it is assigned the same value as any preceding expression that has the same referent.

```
<REFEXP ID="100" REFINDEX="49">IBM</REFEXP> announced
  <REFEXP ID="101"  REFINDEX="50" >a dividend.</REFEXP>
<REFEXP ID="102" REFINDEX="49">It</REFEXP> showed
  <REFEXP ID="103" REFINDEX="51">a hefty fourth quarter profit.
</REFEXP>
```

One of the group participants, Laurent Romary, had recently spent some time investigating the current TEI guidelines and had yet another alternative which made use of empty elements.

A notation that the group used during the workshop discussion used an empty "LINK" element to record all of the link attributes discussed in the preceding section. The following notation essentially captures the information agreed upon, but will undoubtedly undergo further development. Note that the "TYPE" attribute here represents the semantic type of the referent, not the TYPE of link as in the MUCCS annotation shown above.

```
<MARKABLE ID="100" CAT="PN" TYPE="individual">IBM</MARKABLE>
  announced <MARKABLE ID="101" CAT="INDEF-NP" TYPE="individual">
  a dividend.</MARKABLE>
<MARKABLE ID="102" CAT="PRO" TYPE="individual">It</MARKABLE> showed
  <MARKABLE ID="103" CAT="INDEF-NP" TYPE="individual">a hefty
    fourth quarter profit.</MARKABLE>
<LINK LINKTYPE="IDENT" RELATION="ANAPH" SO="100" TA="102">
```

2.5 Current and pending work

During the last meeting of the coreference group, it was agreed that the group should do some post-workshop homework in order to evaluate the proposals that were arrived at during the workshop. We will select short samples from English, French and German. Each participant will code these samples, conforming to his or her understanding of the decisions made at the workshop. We will evaluate the results and develop a set of precise questions to help direct compilation of a first draft of a DRI reference coding manual.

2.6 Future work/open issues

The issues agreed upon as open issues to discuss in the near future are:

- Fluents (entities that change over time)
- Metonymy
- Dialog issues (e.g., two speakers having different interpretations)
- Referential indeterminacy or ambiguity

References

[Chafe 1980] Chafe, Wallace L. 1980. The Pear Stories: Cognitive, Cultural and Linguistic Aspects of Narrative Production. Norwood, NJ: Ablex Publishing Corporation.

[Grishman et al. 1994] Grishman, Ralph; Macleod, Catherine; Meyers, Adam. 1994. COMLEX Syntax: Building a Computational Lexicon Proceedings of COLING-94. Kyoto, Japan.

[Miller et al. 1990] Miller, George A. and Beckwith, Richard and Fellbaum, Christiane and Gross, Derek and Miller, Katherine J. 1990. Introduction to

WordNet: An On-Line Lexical Database. International Journal of Lexicography (special issue) 3(4):235-312.

[Vilain et al. 1995] Vilain, Marc; Burger, John; Aberdeen, John; Connolly, Dennis; Hirschman, Lynette. 1995. A Model-Theoretic Coreference Scoring Scheme. Proceedings of the 6th Message Understanding Conference, pp. 45-52. San Francisco: Morgan Kaufman.

3 Summary of the Multi-Party Dialogue Acts Sub-Group on Forward Looking Communicative Function

David Traum, University of Maryland

3.1 Group Membership

Jens Allwood	jens@ling.gu.se
Toine Andernach	andernac@cs.utwente.nl
Morena Danieli	Morena.Danieli@cselt.stet.it
Barbara Di Eugenio	dieugeni@cs.pitt.edu
Anton Nijholt	anijholt@cs.utwente.nl
Birte Schmitz	birte@cs.tu-berlin.de
Adelheit Stein	stein@darmstadt.gmd.de
David Traum	traum@cs.umd.edu
Syun Tutiya	tutiya@kenon.ipc.chiba-u.ac.jp
Hans de Vreught	J.P.M.deVreught@cs.tudelft.nl

3.2 Issues Resolved during the meeting

3.2.1 Coding Scheme

Issue: many of us were unhappy with some of the names, explanations, distinctions, and coding principles in the Damsl coding manual of Dec. 10th.

Solution: We now have a revised set of distinctions, principles and names, as shown below. We have some preliminary decision trees, indicating how to mark these, as discussed in section 3.3.

Forward Looking Communicative Functions:

- Statement (commitment to belief)
  - Assert
  - Reassert
  - Other
- Addressee's future action
  - "open option" (conditional on addressee's will)
  - Directive
    - Information-Request

        * Action-Directive

- Speaker's commitment to future intention to act
  - "offer" (conditional on addressee's agreement)
  - commit (unconditional commitment)
- Other Specific Function
  - Explicit Performative
  - Exclamations
  - Conventional Openings
  - Conventional Closings
- None

Comments: We still need to code some dialogues to be certain we can use this scheme. This list includes some changes made since the meeting. The first is that the former "conventional" category, which had been split into more logical classes, is now combined under this "other" label. This was to make the main level shorter and easier to organize by menus in a tool. It does not imply in any way that these categories form a natural class. Since each of these are non-hierarchical and relatively rare, compared to the other level functions, it's probably not a problem. The "suggestion" category has been renamed "open option". See the discussion in the suggestion area below.

Some sub-issues we considered in forming the new scheme are:

a) suggestion was missing from the damsl manual. "directive" was meant to cover it, but really suggestions are not always attempts to get the addressee to do something, but often just putting forth possible actions without any obligation imposed by the speaker to seriously consider doing the suggested action. Moreover, many utterances in this category were problematic, systematically confusing inform and directive. e.g., in the trains dialogue 92a-5.1 utt26 from the homework, which talks about future action without trying (hard) to convince the addressee. Our new category is now called "open option". The problem with calling it a "suggestion" is that it doesn't line up well with the performative using the speech act verb "suggest", nor does it include the weak notion of speaker's desire that the addressee does something. Instead this category presents options to the addressee that the speaker doesn't necessarily endorse. This will often appear in conjunction with an statement, where the purpose is really to introduce a possible action option (such as choosing a piece of furniture in the Coconuts domain or an engine in the TRAINS domain or a meeting date in Verbmobil) without the speaker actually wanting the addressee to choose it. This will often be co-marked as a statement if the speaker is also making a claim as well as mentioning an action option.

b) evaluation: These utterances, in which the speaker gives an opinion or reaction to something, were ambiguous between conventional and inform in the old scheme. We decided this was a backward looking function. At the forward level, if the form of the evaluation is some sort of emotive word like "cool!" then it would be marked as expressive. If it is making a claim (e.g., "I think that'll work") it should be marked statement. The backwards group decided not to try to code these, beyond the objective dimension of accepting, rejecting, or "holding" an object on the table for further negotiation. Of course specific coding task designers may add an extra dimension for the emotional content of the backward function.

c) inform - the description was much too narrow, and the name was even more narrow. "Inform" has connotations that the speaker is giving the addressee correct information (that the addressee did not have before). There are many cases in which the speaker talks about the world without actually believing what she says, or without trying to change the belief of the addressee. Our solution was to have a top level act "statement" which is any claim about the world, with subcategory "assert" which is trying to change the belief of the addressee. We did not actually go further to consider cases in which the speaker also believed what he was saying, which would be an inform - this is of course an option that individual task designers could adopt. If the claim has already been made, we add a category "reassert", which also might be a combination of statement and "old" if that is kept in the old/new dimension. This category can also be used for statements in a confirmation/pre-preclosing phase, in which speakers try to verify past agreements, not bringing up new information or trying to change each other's attitudes.

d) the word "commitment" was too broad in scope, since every utterance commits the speaker to some beliefs of courses of action. What we are interested in here is actually commitment to action. After a lot of discussion we ended up with the same basic scheme, though perhaps we need more work on the names. Something coming from this discussion was that statements also include a commitment, in this case a commitment to believing the claim.

e) conventional - this was a bad choice of name, since just about every illocutionary act is conventional. Also, it collapsed together several things which did not belong. We split this class into openings, closings, exclamations, and explicit performatives. It is rare that any of these will be mistaken for another, and trying to lump them together as "conventional" utterances leaves too much temptation to see other things as conventional. One item that is not covered under the new scheme is the kind of "filler" or topic transition "okay" or "(al)right" markers that were analyzed in the damsl manual as conventionally holding the turn. The damsl analysis was flawed, first because there is no convention that these words (as opposed to any other non turn-releasing words) hold the turn, and secondly, the topic management function was not being captured. This kind of function should be covered by a layer concerned with topic management, which might be the information status, or might be something related to higher level structure. Another option is to extend openings and closings to also include topic openings and closings. For now, we do not mark these utterances as having a forward looking function (except perhaps as "exclamations" that everything is "okay" or is "all right"?)

3.2.2 Scope of Illocutionary Acts and This Group

Issue: Problematic nature of the term "illocutionary acts", and conflation of several types of function. The issue was that there are a number of kinds of forces that occur in an utterance that might reasonably termed "illocutionary". Many of these are very similar to the relations to previous utterances that the "response" group was considering.

Solution: Inspired by Allwood's scheme of separating function into "expressive" and "evocative" functions (the former of which expresses attitudes of the speaker about the past and present, and the latter of which "evokes" a future response in the addressee), we decided to consider just the non-responsive functions, that project future (and current) function. Thus, acknowledgments and other feedback (including evaluations of proposals) are not given any function specifically at this layer (unless they also have other functions). The new name of this layer is "forward looking communicative function" (should we abbreviate this flcf?). In parallel, we changed the name of the response group to "backward looking communicative function".

Comments: There are still some difficult elements here. First, this distinction does not quite map directly to Allwood's evocative and expressive categories. This, understandably leads him to some discomfort with the kinds of distinctions we make here. For instance, a statement is a current claim by the speaker, and a commissive commits the speaker at the present moment. Thus there are aspects of Allwood's expressive category at this level. Also, some aspects of the evocative function are not captured here, such as for acknowledgments, the fact that the speaker wants the addressee to realize that the speaker has understood the addressee's prior contribution. I think that moving completely to Allwood's scheme is a more drastic step than the rest of us want to take at this point, but it would be a good (and important) exercise to see if the expressive and evocative functions he describes can be mapped to the sum of forward and backward looking functions that we have in the revised scheme. Likewise, we should go back to the coding schemes that led to Damsl (e.g., Maptask, Verbmobil, Condon) and others (e.g., Delft) to see if we still have problems that can't be overcome by adding new information or making further distinctions (or unions) to what we have here.

3.2.3 How Many Acts Can be Labeled on One Utterance?

Issue: There are really two questions here: can an act be labeled with no functions, and can an act be labeled with more than one function. The difficulty is that allowing multiple labels may make reliability more difficult to achieve - e.g., even though many coders agree on the primary function, some either ignore or disagree about secondary or tertiary function. On the other hand, some utterances really DO perform multiple functions, and trying to decide which is most important can be hard, and might be unnecessary to resolve.

Solution: We decided to include an "unspecified" category when coders felt that there was no forward looking function that was being performed. On the issue of multiple tags, we decided to allow them in principle, but leave the actual decision to the designer of a particular coding task. This would be a parameter that could be set by a coding tool, and should be clearly indicated on any coded data which option was selected.

Comments: Some felt that the unspecified label (no forward communicative function) would never be used. Others felt that, e.g., a pure acknowledgment might have such a label. This dichotomy is probably due to the issues discussed in (2), above. One serious implication of allowing multiple labels is that we must have a set of decision trees rather than one decision tree. This also puts a higher burden on the task designer who wants to disallow multiple tags, since principles for resolving conflicts must be chosen. These, as well, should be clearly marked on the tagged corpus. Some principles include: (a) code high - if an utterance has both "A" and "B" functions, always choose "A" (b) code most important - decide which one is most prominent for this utterance in context.

3.2.4 Lookahead and Intention vs. effect

Issue: There were two related issues here: the first was whether the primary judgment of the coder on these forward looking functions should be based on an assessment of the intention of the speaker or the effect of the utterance on the conversation. While these will mostly be the same (when the speaker successfully communicates his intention), there will be some cases of divergence, as in a misunderstanding. The second issue is how much of the dialogue is the coder allowed to look at to make this determination. While some context (such as the previous few utterances) is important to make sense of the utterance, the actual effects and determinations of intentions might seem different after a few more utterances.

Solution: We code on the actual effect rather than the speaker's intention, since we are only looking from the outside. Of course some determinations will be equivalent to saying the speaker has certain mental states, but we concentrate on the mental states that the speaker is responsible for, not those she actually holds. On the issue of lookahead, this is a parameter that is to be left to the coding task designer. A tagged corpus should clearly state whether the coders were allowed to look ahead (and by how much) when coding an utterance. Also, tools should allow one to set the amount of lookahead allowed.

Comments: It is still arguable how well one can make the distinctions, particularly on actual effect without some amount of lookahead.

3.2.5 Collaborative Completions

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

A: A train will arrive at
B: 4pm

The problem is how to mark this? A's utterance does not have a full force - e.g., he is not actually making a claim about the world. Note that this example is not meant to be a case of a question, where A is eliciting the information from B (as with a final lengthening in English), but a case where A is just not producing the full utterance (in time). The problem is that the claim about the train's arrival is really a collaboration between A and B.

Solution: This issue has not been completely resolved, since it affects both the forward and backward looking groups. Our group proposed two solutions:

(i) Group both utterances into a single segment and code this for the illocutionary force. A problem with this solution is that some are uncomfortable with contributions which cross speaker/turn boundaries.

(ii) Mark each utterance for the full forward looking force, but also provide two special markers at the communicative status level (along with uninterpretable, self-talk, and abandoned): "start-fragment" for the first one, and "continue" for the latter. The advantage is that the second label could also be used when the first part is complete, but some new information is added on, as in a specification or conjunction. The disadvantages are that this "continue" label is in some ways a backward looking function and should be handled at that level, and that this increases the "weird" labels, which gives more overhead to the coding task.

The backward looking group also proposed a third solution: (iii) just mark B's utterance with the backward looking "completion" tag, and A's utterance with it's normal force. Incomplete utterances can be determined by checking forward to responses which are marked completion. A problem is that one would have to look ahead to extract the information. Also, certain cases might not be covered, for example, when the force is completed without a grammatical completion.

Comments: It was decided at the joint forward/backward meeting to isolate some corpus examples of these collaborative completions to see what's really involved with tagging them. 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.

3.2.6 Joint Action

Issue: this concerns talk about future action to be performed by both the speaker and addressee, or more generally negotiations. These acts don't fit neatly into our scheme because the speaker both (contingently) commits to her own action and directs the addressee's action.

Solution: Our tentative solution is to mark these acts with two acts, one indicating speaker's offer/commit, and another for the addressee's future action.

Comments: there was some feeling that this did not really solve the problem, as it cannot distinguish between multiple individual actions and actual joint action. Other solutions include having special markings for joint action. One way would be to segment the proposal and acceptance together and mark it with a joint commitment label. Another would be to have separate labels for joint proposals and acceptance of action. A third would be to parameterize the forward looking layer, so that one might mark the actor (as well as well as other features such as attitude), giving an option of "joint".

3.3 Decision Trees for Coding Scheme

These algol-style decision are from Barbara Di Eugenio, based on my notes — hopefully these are not too controversial, since we agreed on the first two, and the commitment category is not very different from what we had before, except for the realization that commitment/obligation is a wider phenomenon than is covered here.

1. Statement

```
IF S makes a claim C about the world
THEN IF S is trying to change Addressee's beliefs about C
     THEN Assert
     ELSE IF the claim C has already been made
          THEN Reassert
          ELSE Other
ELSE Not a statement
```

2. Addressee's future action

```
IF S is discussing potential action of Addressee
THEN IF S is trying to get Addressee to do something
     THEN Directive
          IF Addressee is supposed to give information
          THEN Information-Request
          ELSE Action-directive
     ELSE (only suggested possibility)
          "open option"
ELSE utterance does not concern Addressee's future action
```

NB: the previous tree doesn't account for cases in which the labels "information-request" and "action-directive" should co-occur, as in information retrieval kind of dialogues

3. Speaker's commitment to future intention to act

```
IF S is potentially committing to intend to perform a future action
THEN IF the commitment is contingent on  Addressee's agreement
     THEN "Offer"
     ELSE unconditional commit
ELSE not about S's commitment to future actions
```

4. Explicit Performatives.

```
IF S is performing an action (not mentioned above, like
        asserting or promising) that is done in virtue of the
        utterance itself (e.g., "I apologize", "thank you")
THEN "Explicit Performative"
```

5. Exclamations

```
IF S utters an exclamation (e.g., "sorry", "ouch!")
THEN "Exclamation"
```

6. Conventional Openings

```
IF S utters a phrase used by convention to summon the
        addressee and/or start the interaction (e.g., "hi",
        "can I help you")
THEN "Conventional Opening"
```

7. Conventional Closings

```
IF S utters a phrase used by convention to dismiss
        addressee and/or (start to) end the interaction
        (e.g., okay or thanking to signal nothing else to do,
        "goodbye")
THEN "Conventional Closing
```

3.4 Current Work

There are several things we need to do in the relatively short term to wrap up this group's mission. These include:

1. carefully going over the whole scheme to see if it fits together
2. coding some dialogues to assist in (1)
3. coordinating with other aspects of the whole scheme - like solving the completion problem with the backward group (by finding and examining examples), and seeing what the implications of functional segmentation and our scheme are with each other, as well as the range of given/new questions.
4. writing up the manual section, with more elaborate text, perhaps revising the decision trees, and adding examples, so that others can use the distinctions we have here.
5. examining proposals on joint action to see if our solutions fall short.
6. non-English examples - we don't want this to be a coding scheme that is only good for English. it would also be helpful for non-readers of these languages if these examples were also given English glosses - perhaps word by word as well as meaning, when possible.
7. Comparisons to existing schemes - one of the main activities at the Penn workshop which was used to construct the Damsl coding scheme was to compare each of the coding schemes for dialogue acts we had used in the pre-workshop homeworks to the new scheme, showing the correspondence. We should also do that for the new scheme, both to those schemes (Verbmobil, Maptask, etc.) and to others that are being used (Gothenburg, Delft, etc.). Can we cover those schemes by adding or merging features or clusters of features?

3.5 Future Work/Open Issues

These are things that we haven't gotten very far or started yet, but which we think are worth doing, perhaps by email, or perhaps in a future meeting.

- Obligation - there are a number of areas in which obligation is important for the acts included in this coding scheme. This occurs not just for commitments to act, but also for statements, in which the speaker has an obligation to believe the claim, and questions where the addressee has some obligations to respond. Also, there are more general obligations that are part of engaging in dialogue. How can we address this issue more fully/consistently?

- Joint Actions - is there a better way for these proposals for joint action (and acceptances) than marking two codings all the time? Should these be special kinds of acts?

- Orthogonality: related to the above two points, is there a better way to extract a set of features from the coding scheme we have that can be marked in parallel to concisely represent the actual function? The backward group seems to have gone in this direction. Also we are fairly close by having separate trees, and allowing multiple codings.

- Feedback: how do we capture the "forward-looking" function of feedback - that is to evoke a response in the addressee that the speaker has (or hasn't) understood what was previously said.

- Self-repair, change, and hesitations - these prominent features of spontaneous (particularly spoken) dialogue do have functions, but are not being covered here. Where do they get covered? This is an important issue for the functional segmentation.

- (new) What to do about confirmation subdialogues? Here the speakers are reviewing plans that have been decided upon, and thus are not trying to get the addressee to do something. What they are doing is mentioning future actions in hope of making things clear. How do we distinguish these mentions of actions and statements from the actual negotiations that established them?

4 Summary of the Response Relations Subgroup on Backward Looking Communicative Function

Johanna Moore, University of Pittsburgh

4.1 Group Membership

James Allen	james@cs.rochester.edu
Morena Danieli	danieli@cselt.stet.it
Peter Heeman	heeman@lannion.cnet.fr
Masato Ishizaki	ishizaki@itl.atr.co.jp
Susanne Jekat	jekat@informatik.uni-hamburg.de
Lori Levin	lsl@cs.cmu.edu
Ian Lewin	ian@cam.sri.com
Diane Litman	litman@cs.columbia.edu
Mario Mast	mast@heidelbg.ibm.com
Johanna Moore	jmoore@cs.pitt.edu
Carol Van Ess-Dykema	cjvanes@afterlife.ncsc.mil
Robert van Vark	R.J.vanVark@cs.tudelft.nl

4.2 Introduction

The group felt that there were three main issues that had to be considered.

1. **What is being responded to?** First, 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. The group consensus was that this issue depends on both the higher-level discourse segmentation and on decisions about the minimal unit of analysis (as discussed in the low-level segmentation group).

   During the discussions, several proposals were made, but we decided to table this part of the discussion in order to make progress on the second item. 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?** Some group members had tried to apply Damsl to other corpora and 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 segmentation issue, collaboration with that group would be necessary to make further progress

on this issue. Therefore, we decided to move on to response types, with the understanding that we expect further work on higher level discourse segmentation to be continued and eventually integrated with our work.

3. **What type of response is given?** Prior to the meeting, although there seemed to be reasonable agreement on the coding task, many people were dissatisfied with the set of response relations. Reasons for dissatisfaction included:

   - Damsl relations conflated what some perceived as different dimensions of the response.
   - There were cases in which coders wanted to use multiple tags for a response relation.

   At the meeting, we focused our efforts on addressing these problems.

4.3 The 4 Dimensions of Response Type

In cooperation with the other groups, we adopted notions from Allwood's scheme of separating function into the **expressive**, which expresses attitudes of the speaker about the past and present, and the **evocative**, which "evokes" a future response in the addressee. Our group's charge was to focus on what the current utterance expresses about the prior discourse. The forward-looking group is attending to those aspects of the utterance that project future (and current) function.

The working group then proposed that there are actually 4 dimensions that must be considered when labeling a response.

1. **Understanding:** This dimension codes for indications of the speaker's ability to recover the semantic content of the utterance.

2. **Agreement:** In this dimension, participants are concerned with coordinating primarily task-level actions. Tags here encode what the utterance indicates about the speaker's attitude toward an action, plan, object, etc.

3. **Informational Relations:** This dimension indicates relationships between the content of the current utterance and the utterance it responds to.

4. **Answer:** This dimension requires more work. It is not clear how to define this dimension, its relationship to information relations, etc. However, many in the group felt we needed such a dimension.

4.3.1 Understanding

This dimension deals with what the current utterance says about the participants ability to recover the explicit content of the antecedent.

The categories are:

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

   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. We had a long discussion about how to deal with (collaborative) completions at the combined forward/backward meeting. These are when one speaker starts to perform an illocutionary function, and it is completed by another speaker. Prototypical examples are:

   ```
   A:  A train will arrive at ...
   B:  4pm
   ```

   The backward-looking group suggests that we mark B's utterance with the backward-looking COMPLETION tag, and that A's utterance be marked with its normal force, here either an ASSERT or REQUEST-INFO depending on whether intonation indicates that this is uttered as the beginning of a statement or as a question. Incomplete utterances can be determined by checking forward to responses that are marked completion. Note that this requires that coders be allowed to look ahead to determine how to label A's utterance. (See David Traum's report from the forward-looking group for other problems and the alternative proposals for handling this issue.

   Other examples:

   ```
   A:  The train arrives at # .... #
   B:  # 4pm on #
   A:  Friday
   ```

   As David Traum notes in the report from the forward-looking group, it was decided at the joint forward/backward meeting to isolate examples of these collaborative completions from out corpora to see what's really involved with tagging them. 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.

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

```
A. Let's take engine E2.
B. You mean E1.
A. E1 to Dansville.
```

* What does it mean to recover?

There was a long discussion at the meeting 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:

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:

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

   sounds fine, whereas

   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 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 "cospecification". Criterion 2 seems to have a simple test, but we will have to see how reliable the coding is on future homeworks.

4.3.2 Agreement

Along the agreement dimension, we identified the following categories:

- **Closing Acts**
  - **Accept**: e.g., "yes".
  - **Accept part**: e.g.,

    ```
    A:  I've got a blue table and a green lamp.
    B:  The table sounds good.
    ```

  - **Non-accept**: e.g, "maybe".
  - **Reject part**:

    ```
    A:  I've got a blue table for $100 and a green lamp for $75.
    B:  The lamp is too expensive.
    ```

  - **Reject**: e.g., "no".
- **Non-closing Acts**
  - **Hold**: this indicates that the topic/issue/alternative that is expressed in the antecedent is still under consideration in some form. This category will often be used when multiple options are being put on the table. When a speaker suggests a different option, it is sometimes a rejection of other options and is sometimes simply another option. Consider the following examples:

A	Can you meet at 7?	
B_1	Well, what about 8?	
		Adds an option
B_2	7 am?	
		Request for clarification of option
B_3	Can we do it at my office?	
		Need further information before making a decision
B_4	Only if it's at my office.	
		Adds a constraint on acceptance

Our analysis of B_1 is that it provides another option without ruling out meeting at 7. We can imagine discourse contexts in which B_1 would be coded as a rejection of A.

4.4 Informational Relations

Many researchers want to specify how the content in utterances is related. Examples are the following:

```
(1)  A.  It's raining outside
     B.  Oh, so we can't go hiking.

         [ACCEPT]  {Consequence}
```

Here the "oh" signals acceptance of A's utterance, and "so we can't go hiking" offers a consequence of A's utterance.

```
(2)  A.  We've got to finish this manual.
     B.  We'll have to work on it tonight.

         [ACCEPT]  {Consequence}
```

This example shows that acceptance does not have to be explicitly marked. Here, B's stating a consequence of A's utterance implicitly indicates acceptance of it.

Note: The set of relations to be used remains to be defined. This is one of the main areas for future work. Research groups should develop sets for their own purposes, and continued work at subsequent DRI meetings is needed to compare, synthesize, and to hopefully come up with a community wide set (possibly at a fairly abstract level).

4.5 Answer

The clear case of this category is in response to questions. Are there other cases?

Problematic Example:

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

(4)  A.  I don't know what time it is.
     B.  It's 3 o'clock.          [ANSWER]

(5)  A.  What time is it?
     B.  I'm not wearing a watch. [REJECT]
```

The group agrees that 3B should not be coded as an answer. Things to note about 3:

- In 3, B is not being cooperative (in the Gricean sense).
- In 3, A did not request information.

4.6 Interaction between backward and forward function

There seems to be general agreement that many benefits come from the clean separation of the forward and backward looking communicative functions of utterances.

The following example shows how a simple dialogue would be coded for both backward and forward-looking communicative function.

```
                                      BACKWARD  |  FORWARD
                                                |
A1.  I have a blue sofa for \$200.        --    | Assert
                                                | Open Option
                                                |
B1.  I have a red one for \$150.          HOLD  | Assert
                                                | Open Option
                                                |
A2.  I don't like red.                  REJECT  | Assert
                                                |
                                                |
B2.  OK, we'll use the blue one.   ACCEPT (A1)  | Commit
```

* Discussion

Note that we do not have an "alternative" tag in either category. Alternatives will now be coded as various combinations, e.g., REJECT (backward function) + OPEN-OPTION or OFFER or COMMIT (forward function).

Evaluations are also gone as a category, but instead arise out of a combination of forward and backward looking function. E.g.,

```
A.  How does that sound?
        Forward:  Action-Directive + Request-Info
B.  Great.
        Backward: Answer + Accept
```

4.7 Issues for Future Meetings

- Higher level discourse structures.
- Integration of coreference coding with forward and backward communicative function.
- 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.

5 Summary of the Segmentation Subgroup

Elisabeth Maier, DFKI

5.1 Group Membership

Jan Alexandersson	janal@dfki.de
Anton Batliner	batliner@informatik.uni-erlangen.de
Robbert-Jan Beun	rjbeun@ipo.tue.nl
Mark Core	mcore@cs.rochester.edu
Nils Dahlbäck	nilda@ida.liu.se
Hans Dybkjær	dybkjaer@cog.ruc.dk
Norman Fraser	norman@vocalis.com
Arne Jönsson	arnjo@ida.liu.se
Susann LuperFoy	luperfoy@starbase.mitre.org
Elisabeth Maier	maier@dfki.de
Joakim Nivre	nivre@ling.gu.se
Norbert Reithinger	bert@dfki.de

(Remark: this group is identical with the subgroup on Information Level / Information Status)

5.2 Motivation

So far, no reliable syntactic / semantic / pragmatic criteria have been available to determine a unit for spoken material. The goal of this subgroup was, therefore, to determine a set of segmentation rules / guidelines to achieve more reliability in segmentation.

5.3 Point of Departure – Segmentation Homework carried out in the Multiparty Group before the Workshop

To get insights into the specific segmentation problems and to find out about the various types of segmentation behavior we distributed a set of dialogues which had to be segmented by the members of the multiparty group. More details about the data and the results are given in the following sections.

5.3.1 Homework Description

The homework material consists of two short English VERBMOBIL dialogues:

- r126.trl – 10 turns
- r150.trl – 7 turns

For each dialogue a transcription and an audio file were provided.

5.3.2 Segmentation Instructions

Group members were asked to segment according to the following instruction:

> A unit is the amount of material that can be attributed one dialogue act / illocutionary function; insert a special character (@) at every segment boundary.

5.3.3 Results

We received the following feedback from the group members: 24 persons submitted segmented data at all; 1 data set was unreadable; two persons submitted late. Therefore, 22 data sets were taken into account for evaluation.

The evaluation method we used for this study was the kappa coefficient. Kappa is computed as follows:

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

where P(A) represents the probability that the annotators agree, while P(E) stands for the probability that the coders agree by chance. The per chance agreement is determined as

$$P(E) = \sum_{i=1}^{n} p_i^2$$

Using this method as a basis we determined[2]:

- the pairwise agreement for the segmentation task;
- the overall agreement (all annotators) for the segmentation task;
- data clusters on the basis of kappa values, with the clusters being sets of labelers;
- typical segmentation problems.

[2]Many thanks go to Michael Kipp for writing the software to automatically evaluate the data.

Pairwise Agreement of Segmentation

	alex	allw	ande	batl	dahl	dani	deVr	dybk	gric	heem	ianl	jons	litm	maie	mcor	nivr	reit	schm	stei	trau	vanV
alexandersson	-	-0.01	-0.04	0.92	0.81	-0.02	0.56	0.75	0.87	0.84	-0.02	0.89	0.89	0.88	-0.06	0.53	-0.04	-0.03	0.79	0.82	0.56
allwood	-	-	0.55	0.00	0.00	0.84	-0.03	0.03	-0.00	-0.00	0.83	0.00	-0.02	-0.00	0.58	-0.02	0.83	0.52	0.03	-0.03	-0.04
andernach	-	-	-	-0.06	-0.05	0.61	-0.09	-0.01	-0.06	-0.04	0.61	-0.06	-0.06	-0.06	0.81	-0.07	0.65	0.86	-0.01	-0.07	-0.08
batliner	-	-	-	-	0.75	-0.01	0.56	0.72	0.86	0.83	-0.01	0.88	0.84	0.87	-0.06	0.45	-0.03	-0.04	0.73	0.79	0.50
dahlbaeck	-	-	-	-	-	-0.01	0.48	0.77	0.75	0.82	-0.02	0.77	0.73	0.74	-0.06	0.55	-0.04	-0.03	0.85	0.72	0.47
danieli	-	-	-	-	-	-	-0.03	0.02	-0.01	-0.01	0.80	-0.00	-0.03	-0.01	0.67	-0.03	0.83	0.63	0.02	-0.04	-0.05
deVreught	-	-	-	-	-	-	-	0.28	0.69	0.61	-0.02	0.61	0.62	0.65	-0.09	0.76	-0.04	-0.09	0.36	0.66	0.80
dybkjaer	-	-	-	-	-	-	-	-	0.67	0.73	0.01	0.69	0.67	0.68	-0.02	0.38	-0.00	0.00	0.91	0.67	0.28
grice	-	-	-	-	-	-	-	-	-	0.88	-0.01	0.91	0.87	0.92	-0.06	0.55	-0.03	-0.04	0.74	0.88	0.65
heeman	-	-	-	-	-	-	-	-	-	-	-0.01	0.88	0.86	0.87	-0.05	0.50	-0.03	-0.02	0.77	0.87	0.56
ianlewin	-	-	-	-	-	-	-	-	-	-	-	-0.01	-0.04	-0.01	0.56	-0.01	0.77	0.55	0.01	-0.05	-0.03
jonsson	-	-	-	-	-	-	-	-	-	-	-	-	0.84	0.94	-0.06	0.48	-0.02	-0.03	0.76	0.83	0.58
litman	-	-	-	-	-	-	-	-	-	-	-	-	-	0.90	-0.07	0.49	-0.06	-0.04	0.71	0.86	0.62
maier	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-0.06	0.50	-0.03	-0.04	0.75	0.87	0.60
mcore	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-0.08	0.65	0.76	-0.03	-0.08	-0.09
nivre	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-0.04	-0.08	0.44	0.49	0.73
reithinger	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.64	-0.01	-0.05	-0.05
schmitz	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.00	-0.03	-0.09
stein	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.71	0.35
traum	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.54
vanVark	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Overall Agreement of Segmentation The segmentation agreement between all coders was computed as

kappa(21) = 0.37708938

Clusters on the basis of kappa values With an off-the-shelf clustering algorithm we determined sets of coders that agreed in their segmentation behavior. The clusters were determined on the basis of the kappa values, i.e. coders between who a certain kappa value could be achieved were taken together in one set.

An example for such a cluster set looks as follows:

Cluster 0: n = 9 persons; kappa = 0.8192
Cluster 1: n = 4 persons; kappa = 0.8221
Cluster 2: n = 3 persons; kappa = 0.8149
Cluster 4: n = 3 persons; kappa = 0.8489

Typical Segmentation Problems In the segmented data various sources of disagreement could be identified. They concerned the segmentation of

- particles / conjunctions (*well, okay, alright*)
- hesitations
- coordinated sentences (e.g. linked by *and*)
- subordinate sentences (e.g. *if ... then*)
- reformulations
- suggestions and a request for their confirmation

5.4 Issues Touched Upon in the Meeting

Among a whole set of factors relevant for segmentation we identified the following:

- behavioral / action-oriented
- grammatical
- functional / pragmatic
- prosodic

Some group members pointed out that it has been shown in empirical studies that the segmentation of dialogues changes only slightly when the annotators can listen to speech data. Also, it was noted that the segmentation principles may differ depending on the purposes for which the data are used.

A problem was seen in the segmentation of discontinuous actions which follow multiple purposes. Segmentation may also differ depending on the annotation level being looked at (e.g. illocutionary acts, backward looking, etc.). Segmentation principles and annotation schemes are interdependent and cannot be separated.

Despite all these problems an annotation scheme has to be robust against different segmentation principles.

5.5 Our Solution: Segmentation Types and Segmentation Rules

We developed a set of segmentation types and rules which address and hopefully also solve the segmentation problems identified above:

5.5.1 Segment Types

We distinguish three types of segment boundaries:

- **regular** segment boundaries
- **weak** segment boundaries
- **drop-in** segment boundaries

Regular Segment Boundaries Boundaries of this type are placed around material that serves an illocutionary/communicative function. In the following such boundaries are marked by '@'.

Example:

well @ why don't we just met on twenty third @ let us get an early start @ at nine sounds good @

Weak Segment Boundaries Weak segment boundaries (indicated in the following by '*') can be used to subsegment regular segments into smaller units, e.g. they may be used to segment a multi-sentence unit into syntactical units (see example below). The resulting units belong to the same illocutionary type. The marking of weak boundaries is optional and will be used by groups that need data which are segmented in a rather fine-grained fashion.

Example:

*I am afraid that I am out of town next week * and I come back on Thursday * but I am tied up in the morning @ Friday I have completely free * that would be Friday the twenty third @*

Drop-In Segment Boundaries Drop-in segment boundaries are used to mark phenomena like self-repair and hesitations. In the following drop-in boundaries are marked by $.

Example:

Let's meet on $ Tuesday, I mean $ Wednesday @

By distinguishing three types of segment boundaries we allow for nested segmentation. Also, the three boundary types give rise to different possible uses for the various segment types: weak segment boundaries, for example, can be used to introduce segments of smaller size. These boundaries can be ignored should they not be needed by a given application. Drop-in boundaries mark phenomena that are irrelevant for specific applications.

5.5.2 Segmentation Rules

1. segment material that serves an illocutionary function (@).
2. when in doubt whether to segment or not, **don't** segment.
3. if there are strong indicators, e.g. prosodic markers like a long pause, segment (@). (note: segment only in cases that are compatible with rule 1.)
4. in collaborative completions: segment at locations of speaker change (@).
5. optional: subsegment material into smaller units using weak boundaries (*) where the resulting units serve the **same** illocutionary function.
6. for drop-in phenomena like self-repair and hesitations: introduce $-boundaries.

5.6 Issues Delegated to other Working Groups

When attributing illocutionary functions to drop-in phenomena the segmentation rule 6. is not compatible with segmentation rule 1. To this end we asked the "Forward-Looking Group" to consider the introduction of Illocutionary Acts that cover phenomena like self-repair. Since segmentation depends highly on the illocutionary functions (see rule 1.) close cooperation with the "Forward-Looking Group" is required.

5.7 Future Actions

In future meetings the following points need further discussion:

- **Segmentation and Prosody**
  We need to examine how prosody contributes to the determination of

segment boundaries. To this end David Traum volunteered to organize a homework in preparation of the next meeting.

- **Data representation (e.g. SGML)**
  We need to discuss whether SGML is the right markup language for segmentation and markup. We need to have a look at the shortcomings of SGML and find a way to circumvent them.

- **Stability of segmentation principles across languages**
  We need to make sure that the segmentation principles specified in this meeting also work for languages other than European languages.

- **Notational problems of segmentation**
  We have to clarify where segment boundaries have to be placed (at the beginning AND the send of a segment or only at the beginning?) We have to find a way to minimize notational problems, e.g. by indexing boundaries.

6 Summary of the Information Level and Information Status Subgroup

Norbert Reithinger, DFKI[3]

6.1 Group Membership

Jan Alexandersson	janal@dfki.de
Anton Batliner	batliner@informatik.uni-erlangen.de
Robbert-Jan Beun	rjbeun@ipo.tue.nl
Mark Core	mcore@cs.rochester.edu
Nils Dahlbäck	nilda@ida.liu.se
Hans Dybkjær	dybkjaer@cog.ruc.dk
Norman Fraser	norman@vocalis.com
Arne Jönsson	arnjo@ida.liu.se
Susann LuperFoy	luperfoy@starbase.mitre.org
Elisabeth Maier	maier@dfki.de
Joakim Nivre	nivre@ling.gu.se
Norbert Reithinger	bert@dfki.de

6.2 Information Level

The Damsl Manual proposed five categories for the annotation of the information level:

- Task
- About-Task
- Communication
- About-Communication
- Non-Relevant

During the discussions we had some problems with these classes

- dialogs have a global task, which has subtasks, one (some) of which are communicative tasks. Therefore, all Communication related classes can be folded into Task related.
- For some purposes you need a more fine grained distinction than task/about task, e.g.

[3]Based on the summary slides of Susan LuperFoy, MITRE

  - problem-solving process
  - domain

We also made the observation that at some level every utterance is about the task. However, some utterances advance progress towards the task and others foster communication

Example:

Order: *"Hand me the grommet"*

Reply:

- *"What's a grommet"*
- *"Which grommet"*
- *"Did you say 'grommet'"*
- *"How do you spell grommet"*

The following utterance types can be (almost) always be labelled as communicative

- greetings
- closings
- moves to maintain contact
  - perception
  - turn-taking
  - understanding previous contributions (both the propositional content and the illocutionary act)

Problematic cases are utterances like, e.g. service offers ("May I help you with something else?"). In a computer-based telephony solution this utterance may belong to the task, while in others it may belong to the communicative level.

For some utterances from the homework material we made a vote using the four categories Task (T), About-Task (AT), Communication (C), and About-Communication (AC), allowing multiple votes. The results are shown in the following table:

	T	AT	C	AC	utterance
(1)	14	10	1	0	I'd sure check for you
(2)	3	6	16	0	How can I help you
(3)	7	8	7	3	Let me check
(4)	8	10	3	4	I goofed
(5)	11	0	6	0	Now it is for ...

As can be seen, there was no unanimous vote. For (1), the votes were mainly split between TASK and ABOUT-TASK. In the discussions it became clear, that it depends on the definition of the task or the application area whether it belongs to either one of these categories. For (2) the same argument was made, however the COMMUNICATION got a majority vote. Utterance (3) shows that for utterances that put the listener on a hold while doing some task, it is not clear at all to which class it belongs. While for (4) there is a tendency to classify it as "tasky", (5) clearly shows again that the task and communicative function may be just the two sides of one coin.

The annotation also depends on the perspective of the annotator: is s/he

- the speaker
- the hearer
- an omniscient diety

Another factor is the context that is seen during annotation. Is only the preceding context known, or does the annotator also take the following utterances into account during annotation?

For the coding manual, we recommend the following definition:

> *Communication* – designed to maintain contact, perception, understanding, turn-taking, and previous contributions

We came up with four alternative proposals for the info-level markup (the need for the NON-RELEVANT category depends on the consumer of the scheme):

1. keep the 5-way distinction from the Damsl manual
2. make a three-way distinction # 1: the task includes also all communication related utterances
   - TASK
   - ABOUT-TASK
   - NON-RELEVANT
3. three-way distinction # 2: the ABOUT categories are omitted
   - TASK
   - COMMUNICATION
   - NON-RELEVANT
4. four-way distinction: there is only one category dealing with the communication level
   - TASK
   - ABOUT-TASK

- Communication
- Non-Relevant

We did not come up with a clear decision, which of the four proposals should be selected. However, we favored the second or third solution.

6.3 Information-Status: Old/New

Before defining tags that deal with an old/new distinction one has to consider the difference between utterances to be tagged with the labels and think about possible uses for these tags. Example of uses are

- inference
- determination of focus at the propositional level
  - new
  - elaboration
  - old
- slots of the task getting tracked/which values get revisited
- voice-only in noisy environment

Problems with the old/new distinction arise as

- almost everything can be thought of as new in some sense
- a single utterance may have new and old components
- a single utterance may have components that are new/old at the communication level and old/new at the task level
- in some texts it is hardly useful (as in the Coconut homework)
- the name might be misleading and can be exchanged with ancient/novel, given/new, introduced/not introduced, ...
- the distinction may be based too much on repetition of strings on the surface
- it is either based on (too) shallow features or
- if it comes to a deep functional view, one has to deal with phenomena like grounding, feedback, and intention; this is influenced by the backward looking functions.

Another point is whether the distinction is exhaustive. Especially old can be subdivided further into different types

- repetition $\longrightarrow$ anaphora
- reformulation ($\neq$ paraphrase)
- inference $\longrightarrow$ to bridge anaphora (but not all logically feasible inferences must be drawn)

For the revised version of the coding manual, we considered four possible schemes for the old/new distinction:

- keep old/new
- add irrelevant
- subdivide old as shown above
- define four categories of info status
  - repetition
  - reformulation
  - inference
  - new

Someone who is interested in tracking content of common ground might be interested in watching these tags, and the relationships between these and others.

The general maxim for the selection of the scheme we agreed on was

> *If it doesn't interfere and there's intuition that is might/will be useful, then keep it.*