Lang Res Eval (2006) 40:5–23
DOI 10.1007/s10579-006-9006-4

ORIGINAL PAPER

A corpus for studying addressing behaviour in multi-party dialogues

Natasa Jovanovic · Rieks op den Akker · Anton Nijholt

Published online: 19 October 2006

Abstract This paper describes a multi-modal corpus of hand-annotated meeting dialogues that was designed for studying addressing behaviour in face-to-face conversations. The corpus contains annotated dialogue acts, addressees, adjacency pairs and gaze direction. First, we describe the corpus design where we present the meetings collection, annotation scheme and annotation tools. Then, we present the analysis of the reproducibility and stability of the annotation scheme.

Keywords Addressing · Multi-party dialogues · Multimodal corpora · Annotation schemas · Reliability analysis

1 Introduction

Current tendencies in modelling human–computer as well as human–human interactions are moving from a two-party model to a multi-party model. One of the issues that becomes salient in interactions involving more than two parties is addressing. Addressing as an aspect of every form of communication has been extensively studied by conversational analysts and social psychologists (Clark & Carlson, 1992; Goffman, 1981; Goodwin, 1981). Recently, addressing has received considerable attention in interaction modelling in the context of mixed human–human and human–computer interaction (Bakx, van Turnhout, & Terken, 2003; van Turnhout, Terken, Bakx, & Eggen, 2005), human–human–robot interaction (Katzenmaier, Stiefelhagen, & Schultz, 2004), mixed human-agents and multi-agents interaction (Traum, 2004) and multi-party human–human interaction (Jovanovic & op den Akker, 2004; Otsuka, Takemae, Yamato, & Murase, 2005).

Addressing is carried out through various communication channels, such as speech, gestures or gaze. To explore interaction patterns in addressing behaviour

N. Jovanovic · R. op den Akker (✉) · A. Nijholt
Human Media Interaction Group, University of Twente, PO Box 217, Enschede 7500 AE, The Netherlands
e-mail: infrieks@ewi.utwente.nl

and to develop models for automatic addressee prediction, we need a collection of audio and video interaction recordings that contains a set of annotations relevant to addressing. Meetings as complex interplays of interacting participants represent a relevant domain for the research on different aspects of interactions involving more than two participants who employ a variety of channels to communicate with each other.

In the context of the meeting research, several corpora have already been developed. Some of the existing meeting corpora, such as the ICSI (Janin et al., 2004) and ISL (Burger & Sloane, 2004) corpora—currently widely used to study linguistic phenomena in natural meetings—are limited to audio data only. The NIST audio–visual meeting corpus (Garofolo, Laprun, Michel, Stanford, & Tabassi, 2004) is designed to support the development of audio and video recognition technologies in the context of meetings. Currently, it provides transcriptions of the meetings to enable the research on automatic speech recognition in meetings. To support the research on higher-level meetings understanding, the VACE (Chen et al., 2006) and AMI (Carletta et al., 2006) multi-modal meeting corpora are currently being produced. The VACE corpus is being developed to support research on multimodal cues, such as speech, gaze, gestures and postures, for understanding meetings. The AMI data collection is being developed to enhance research in various areas related to the development of meeting browsing technologies and remote meeting assistants, including speech recognition, computer vision, discourse and dialogue modelling, content abstraction, human–human and human–computer interaction modelling. It contains a range of annotations including, among others, speech transcription, dialogue acts, topic segmentation, focus of attention, head and hand communicative gestures, and summaries.

In this paper, we describe a multi-modal corpus of hand-annotated meeting dialogues, designed for studying addressing behaviour in face-to-face conversations. The meetings were recorded in the IDIAP meeting room in the research program of the European M4[1] and AMI[2] projects. The recordings are available through the MultiModal Media File Server.[3] Currently, the corpus contains hand-annotated dialogue acts, adjacency pairs, addressees and gaze directions of meeting participants. A set of the corpus' annotations of the M4 meetings is available as a part of the M4 meeting collection.[4]

Apart from the corpus description which includes the description of the meeting data, annotation scheme, annotation tools and the corpus format, this paper reports the reliability of the overall annotation scheme as well as a detailed analysis of detected sources of unreliability.

2 Meeting data

The corpus consists of 12 meetings recorded at the IDIAP smart meeting room (Moore, 2002). The room is equipped with fully synchronized multi-channel audio

[1] The M4 (MultiModal Meeting Manager) project: http://www.m4project.org

[2] The AMI (Augmented Multi-party Interaction) project: http://www.amiproject.org

[3] MMM File Server http: //www.mmm.idiap.ch

[4] http://www.mmm.idiap.ch/M4-Corpus/annotations/NXTbasedAnnotation/

and video recording devices (see Fig. 1). Of the 12 meetings, 10 were recorded within the scope of the M4 project. These meetings are scripted in terms of type and schedule of group actions that participants perform in meetings such as presentation, discussion or note taking, but content is natural, spontaneous and unconstrained. Spontaneous behaviour of participants in these meetings allows us to examine observable patterns of addressing behaviour in small group discussions. More natural, scenario-based, meetings have been recorded in the scope of the AMI project. One of the AMI pilot meetings recorded at the IDIAP meeting room is included in our corpus. The meeting involves a group focused on the design of a TV remote control. The last meeting in our corpus is one of a series of meetings recorded at IDIAP for the exploration of argumentative structures in meeting dialogues.

Research on small group discussions presented in (Carletta, Anderson, & Garrod, 2002) has shown that there is a noticeable difference in the interaction patterns between large and small groups. A small group discussion involving up to seven participants resembles two-way conversations that occur between all pairs of participants and every participant can initiate conversation. A large group discussion is more like a series of conversations between a group leader and various individuals with the rest of participants present but silent. In the M4 and AMI data collection each meeting consists of 4 participants. Hence, the meetings in our corpus satisfy the interaction patterns of small group discussions. There are 23 participants in the corpus. The total amount of recorded data is approximately 75 min.

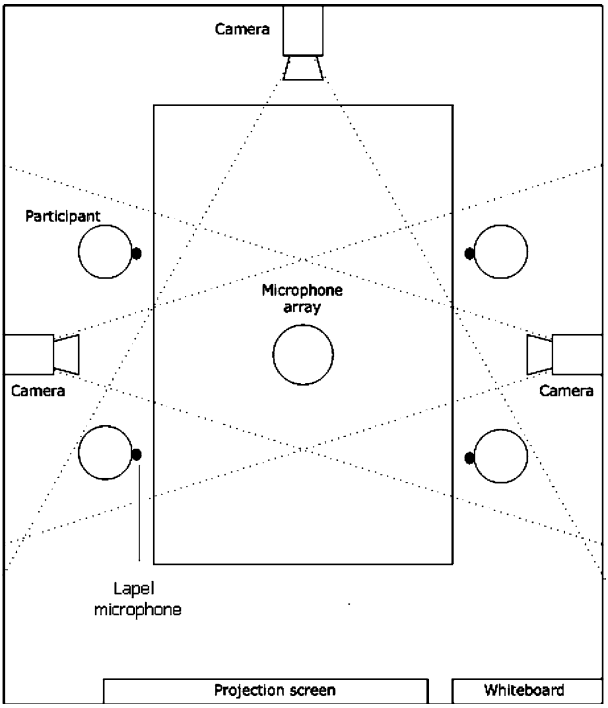

Fig. 1 The configuration of the IDIAP meeting room (M4 data collection)

3 Annotation scheme

In two-person dialogues, it is usually obvious to the non-speaking participant who is the one being addressed by the current speaker. In a multi-party case, the speaker has not only the responsibility to make his speech understandable for the listeners, but also to make clear to whom he is addressing his speech.

Analysis of the mechanisms that people use in identifying their addressees leads to a model of a conversation that describes the features that play a role in these mechanisms. Our annotation scheme is based on the model presented in (Jovanovic & op den Akker, 2004). The features described in the model are of three types: verbal, nonverbal and contextual. For example, utterances that contain the proper name of a conversational participant may be addressed to that participant. Also speaker gaze behaviour may be a feature that gives a hint to the intended addressee. The history of the conversation is important as well, since most of the utterances that are related to the previous discourse are addressed to one of the recent speakers.

Although the model contains a rich set of features that are relevant for observers to identify the participants the speaker is talking to, due to time constraints, the meetings were annotated with a subset of the selected properties. In addition to addressee annotation, the corpus currently contains annotations of dialogue acts, adjacency pairs and gaze direction. We also considered coding of deictic hand gestures as they can be used as a means of addressing. However, it was found that deictic hand gestures occur very rarely in the data.

3.1 Dialogue acts

Annotation of dialogue acts involves two types of activities: marking of dialogue acts segment boundaries and marking of dialogue acts themselves.

Utterances within speech transcripts, also known as prosodic utterances, were segmented in advance using prosody, pause and syntactical information. In our scheme, a dialogue act segment may contain a part of a prosodic utterance, a whole prosodic utterance, or several contiguous prosodic utterances of the same speaker.

Our dialogue act tag set is based on the MRDA (Meeting Recorder Dialogue Act) set (Dhillon, Bhagat, Carvey, & Shriberg, 2004). The MRDA tag set represents a modification of the SWDB-DAMSL tag set (Jurafsky, Shriberg, & Biasca, 1997) for an application to multi-party meeting dialogues. Each functional utterance in MRDA is marked with a label, made up of one or more tags from the set. The analysis of the MRDA tag set presented in (Clark & Popescu-Belis, 2004) shows that the number of possible labels reaches several millions. For that reason, the usage of the complete set may lead to a low quality of manual annotations.

Unlike MRDA, each utterance in our dialogue act annotation scheme is marked as *Unlabelled* or with exactly one tag from the tag set that represents the most specific utterance function. For addressee identification, it is less important whether an utterance is a suggestion in the form of a question or in the form of a statement. More important is that the speaker suggests to the addressee to perform an action, informing all other participants about that suggestion.

Our dialogue act tag set is created by grouping some of the MRDA tags into 17 categories that are divided into seven groups, as follows:

- Statements
- *Statement* [MRDA: Statement]. The Statement tag marks utterances which are objective and factual statements as well as utterances which are opinions and other subjective statements.
- Acknowledgements and Backchannels
- *Acknowledgement* [MRDA: Acknowledgement, Backchannel]. The Acknowledgement tag is a common tag used for acknowledgements and backchannels. Acknowledgements are utterances in which a speaker acknowledges a previous speaker's utterances or a significant portion of a previous speaker's utterance. They are neither positive nor negative. Backchannels have a function to show that a listener is paying attention. They are made in the background by a speaker who does not have the floor.
- *Assessment/Appreciation* [MRDA:Assessment/Appreciation]. The Assessment/Appreciation tag marks utterances that are acknowledgements directed to another speaker's previous utterance with slightly more emotional involvement. They can be positive, such as ''that's great'', ''wow!'', or negative, such as ''not good enough'', ''that's impossible''.
- Questions
- *Information Request* [MRDA: Wh-Question, Y/N Question, OR-Question, Or Clause After Y/N Question]. The Information Request tag marks questions that require specific answers. Examples include ''what kind of preprocessing are you using?'' or ''but do you often cook at night?''
- *Open-ended Question* [MRDA: Open-ended Question]. The Open-ended Question tag marks questions that do not require a specific answer; they are rather asked in a broad sense (e.g., ''What about you?'' or ''anything else?'').
- *Rhetorical Question* [MRDA: Rhetorical Question]. The Rhetorical Question tag marks questions that are used for rhetorical effects. No answer is expected to those questions. Examples include ''who knows?'' or ''who would have thought that it was possible?''
- Responses
- *Positive Response* [MRDA: (Partial) Accept, Affirmative Answer]. The Positive Response tag marks utterances that exhibit a (partial) agreement to or a (partial) acceptance of or an affirmative answer to a previous speaker's proposal, statement or question.
- *Negative Response* [MRDA: (Partial) Reject, Dispreferred Answer, Negative Answer]. The Negative Response tag marks utterances which exhibit a (partial) disagreement to or a (partial) reject of or an explicit or implicit negative answer to a previous speaker's proposal, statement or question.
- *Uncertain Response* [MRDA: Maybe, No Knowledge]. The Uncertain Response tag marks utterances which express a lack of a speaker's knowledge regarding some subject or that a speaker's utterance is probable, yet not definite (e.g., ''maybe'', ''I am not sure'').

- Action Motivators
- *Influencing-listeners-action* [MRDA: Command, Suggestion]. The Influencing-listeners-action tag marks utterances which influence the listeners' communicative or non-communicative future actions such as commands, suggestions, proposals, advices.
- *Committing-speaker-action* [MRDA: Command, Suggestion]. The Committing-speaker-action tag marks utterances which indicate that a speaker has committed himself, in varying degrees of strength, to some future course of action. The speaker can explicitly commit himself that he will execute an action at some point in the future, such as ''I will prepare a presentation for the next meeting'', or he can suggest that he will do so if listeners accept it, such as ''I can say something about that''.
- Checks
- *Follow Me* [MRDA: Follow Me]. The Follow Me tag marks utterances by which a speaker wants to ensure that what he is saying has been understood by listener(s) (e.g., ''do you understand?'', ''okay?'', ''this is clear?'').
- *Repetition Request* [MRDA: Repetition Request]. The Repetition Request tag marks utterances in which a speaker wants another speaker to repeat all or a part of a previous utterance. This is mostly the case when a speaker could not hear or could not interpret what another speaker has said and wants to hear it again.
- *Understanding Check* [MRDA: Understanding Check]. The Understanding Check tag marks utterances in which a speaker wants to make sure whether he understands what a previous speaker said or whether he understands some sort of information. Examples include ''you said that machine learning techniques are applicable?'', ''so this part is new, right?''.
- Politeness Mechanisms
- *Thanks* [MRDA: Thanks]. The Thanks tag marks utterances in which a speaker thanks another speaker(s).
- *Apology* [MRDA: Apology]. The Apology tag marks utterances in which a speaker apologizes for something he did (e.g., coughing, interrupting another speaker) or he plans to do (e.g., to leave meeting earlier, to make a phone call during the meeting).
- *Other polite* [MRDA: Welcome, Downplayer, Sympathy]. The Other polite tag marks all other acts of politeness that do not contribute to the overall discussion but rather have a social impact such as, ''you're welcome'', ''I'm kidding'', ''good luck'', ''success'', ''you are so nice''.

The MRDA scheme also allows the annotation of turn-taking (e.g., floor grabber) and turn-maintaining (e.g., floor holder) mechanisms. The turn managing dimension of utterances' functions is excluded from our scheme. Utterances that function only as turn taking, turn giving or turn holding signals are marked as *Unlabelled*. Turn-taking and addressing as two aspects of conversational interactions are related, but we were specifically interested in studying how addressing actually works, i.e., how people address each other, in order to build predictive models for addressee identification. The scheme also excludes (1) a set of MRDA tags that are related to restating information such as repetitions and corrections, (2) a set of MRDA tags that are related to rhetorical roles such as explanations or elaborations and (3) a set

of MRDA tags that provide further descriptions of utterance functions such as self talk, third party talk, jokes, meeting agendas or topic change.

3.2 Adjacency pairs

Adjacency pairs (APs) are minimal dialogic units that consist of paired utterances such as question–answer or statement–agreement. The paired utterances are produced by different speakers. Utterances in an adjacency pair are ordered with the first part (A-part, the initiative) and the second part (B-part, the response). In multi-party conversations, adjacency pairs do not impose a strict adjacency requirement, since a speaker has more opportunities to insert utterances between two elements of an adjacency pair. For example, a suggestion can be followed by agreements or disagreements from multiple speakers.

In our scheme, adjacency pairs are labelled at a separate level from dialogue acts. Labelling of adjacency pairs consists of marking dialogue acts that occur as their A-part and B-part. If a dialogue act is an A-part with several B-parts, for each of these B-parts, a new adjacency pair is created. Furthermore, each dialogue act is marked as a B-part of at most one and as an A-part of an arbitrary number of adjacency pairs. Although it is theoretically possible that a B-part is related to several A-parts, for example, an utterance may answer two questions, the analysis of the data showed that these cases hardly occur.

3.3 Addressees

In a group discussion, many of the speaker's utterances are addressed to the group as a whole. However, the speaker may show by verbal or non-verbal behaviour that he intends to affect one selected participant or a subgroup of participants in particular, that he expects that participant or that subgroup to react on what he says. In this case, the selected participant or the subgroup is the addressee of the dialogue act performed by the speaker.

Given that each meeting in the corpus consists of four participants, the addressee tag set contains the following values:

- a single participant: P_x
- a subgroup of participants: P_x,P_y
- the whole audience: P_x,P_y,P_z
- *Unknown* $x,y,z \in \{0,1,2,3\}$; P_x denotes speaker at the channel x. The *Unknown* tag is used when the annotator cannot determine to whom the dialogue act is addressed. It is also used for utterances marked as *Unlabelled* and for backchannels.

3.4 Gaze direction

Annotation of gaze direction involves two types of activities: labelling the changes in the gazed targets and labelling the gazed targets themselves.

For addressee identification, the only targets of interest are meeting participants. Therefore, the tag set contains tags that are linked to each participant (P_x) where $x \in \{0,1,2,3\}$ and the *NoTarget* tag that is used when the speaker does not look at any of the participants. The set can be further refined by adding some objects of interests in the meeting room such as whiteboard, projector screen or notebook.

Since the meeting room was not equipped with close-up cameras the gaze information was induced based on the side and central cameras (see Fig. 1). This was the main reasons for not imposing the requirement for a high precision in labelling changes in the gazed targets.

4 Annotation tools

The corpus was created using two annotation tools developed at the University of Twente: the DACoder (Dialogue Act Coder) and the CSL (Continuous Signal Labelling) tools (Reidsma, Hofs, & Jovanovic, 2005). The DACoder supports annotation of dialogue acts, addressees and any kind of relations between dialogue acts such as adjacency pairs or rhetorical relations. The CSL tool supports labelling of time-aligned annotation layers directly related to the signal files. Any annotation layer that consists of simple labelling of non-overlapping segments of the time line can be coded using this tool (e.g., gaze directions or postures).

The tools were built using NXT (NITE XML Toolkit) (Carletta et al., 2003). NXT uses a stand-off XML data storage format which consists of several inter-related xml-files. The structure and location of the files are represented in a ''metadata'' file. The NXT stand-off XML format enables the capture and efficient manipulation of complex hierarchical structures across different modalities. Furthermore, it supports an easy extension of the corpus with new annotation layers without influencing exiting annotations. For exploitation of annotated data, NXT provides the NXT Search tool for the execution of the queries expressed in the NXT Query Language (NQL).[5]

5 Distributional statistics

In this section, we provide distributional statistics for the dialogue act tags and addressee tags in the corpus. The corpus contains 1457 dialogue act segments out of which 131 segments (8.99%) are labelled as *Unlabelled*. Table 1 shows the distribution of DA tags after discarding those segments that are marked as *Unlabelled*.

The distribution of the addressee tags over those segments that are marked with a DA label is presented in Table 2. All subgroup addressee tags (P_x,P_y) are grouped into the SUB category and all tags that denote the whole audience (P_x,P_y,P_z) are grouped into the ALLP category.

6 Reliability

In order to obtain valid research results, data on which they are based must be reliable. We have performed two reliability tests proposed by Krippendorff (1980): stability (intra-annotator reliability) and reproducibility (inter-annotator reliability). Stability is the degree to which an annotator's judgments remain unchanged over time. It is measured by giving the same annotator a set of data to annotate twice, at different times. Reproducibility is the degree to which different annotators can

[5] NXT Query Language: http://www.ims.uni-stuttgart.de/projekte/nite/

Table 1 Distribution of DA tags

Statement	44.19%	Influencing-listeners-action	3.32%
Information request	9.58%	Committing-speaker-action	2.64%
Open-ended question	2.11%	Follow me	0.15%
Rhetorical question	0.60%	Repetition request	0.53%
Acknowledgement	15.61%	Understanding check	1.58%
Assessment/Appreciation	2.19%	Thanks	0.38%
Positive response	11.99%	Apology	0.08%
Negative response	3.24%	Other polite	0.30%
Uncertain response	1.51%		

Table 2 Distribution of addressee tags

Addressee	
P0	12.97%
P1	14.63%
P2	17.50%
P3	16.59%
ALLP	34.16%
SUB	1.73%
Unknown	2.41%

produce the same annotation. It is measured by giving several annotators the same data to annotate independently, following the same coding instructions.

Reliability is a function of agreement achieved among annotators. In the dialogue and discourse processing community, the Kappa agreement coefficient (κ) has been adopted as a standard (Carletta, 1996; Cohen, 1960). In recent years, there have been some discussions about the usage of Kappa as an appropriate reliability metric. Krippendorff's Alpha (α) has been proposed as a more adequate metric for assessing reliability of subjective codings (Krippendorff, 1980, 2004).

To estimate reliability of dialogue act, addressee and gaze annotation, we applied both agreement coefficients. The obtained Kappa and Alpha values were identical. Therefore, in the following sections we report only Kappa values. In contrast to dialogue act and addressee annotation, adjacency pairs annotation cannot be considered as a simple labelling of annotation units with categories. Therefore, we developed our own approach that represents annotated APs in a form of categorical labelling and measures agreement on APs annotation using Alpha.

For the evaluation of Alpha and Kappa values, we used Krippendorff's scale that has been adopted as standard in the discourse and dialogue processing community (Krippendorff, 1980). According to that scale, any variable with an agreement coefficient below .67 is disregarded as unreliable, between .67 and .8 allows drawing tentative conclusions and above .80 allows drawing definite conclusions.

6.1 Detecting sources of unreliability

Detecting causes of disagreement may be of great use to obtain reliable data or to improve data reliability. A source of unreliability can be a coding unit, a category, a subset of categories or an annotator (Krippendorff, 1980). Even if a category is well defined, annotators may still have different interpretations of the category. Furthermore, annotators may show a correlated disagreement. For example, annotator A_1 uses category C_1 whenever annotator A_2 uses category C_2.

To identify which categories are sources of unreliability we measured single-category reliability (Krippendorff, 1980). Single-category reliability assesses the extent to which one category is confused with all other categories in a set. It is estimated by grouping the remaining categories into one category and measuring the agreement among annotators regarding the assignment of units to these two categories. A low agreement can be the result of an ambiguous definition of the category or of the coders' inability to interpret the meaning of the category.

7 Inter-annotator reliability

In this section we present the analysis of inter-annotator reliability of the annotation scheme applied on the M4 meeting data.

Six trained annotators were involved in the corpus creation. They were divided into two groups: the DA (Dialogue Act) group and the VL (Video Labelling) group. The DA group, involving 4 annotators, annotated dialogue acts, addressees and adjacency pairs. The VL group, involving 2 annotators, annotated gaze direction. The corpus was divided into two sets of meetings. The DA group was divided into 2 subgroups of 2 annotators: the B&E group and the M&R group. Each of these subgroups annotated exactly one set of meeting data. Each annotator in the VL group annotated one set of meeting data. Additionally, two meetings were annotated by both annotators in the VL group in order to test reliability of gaze annotation. In summary, each meeting in the corpus was annotated with dialogue acts, addressees and adjacency pairs by exactly two annotators, and with participants' gaze directions by at most two annotators.

7.1 Reliability of dialogue acts annotation

We first measured agreements among annotators on how they segmented dialogues into dialogue act segments. Then, we tested reliability of dialogue act classification on those segments for which annotators agreed on their boundaries.

7.1.1 Segmentation reliability

In the discourse and dialogue community, several approaches have been proposed for assessing segmentation reliability using various metrics: percent agreement (Carletta et al., 1997; Shriberg, Dhillon, Bhagat, Ang, & Carvey, 2004), precision and recall (Passonneau & Litman, 1997), and κ (Carletta et al., 1997; Hirschberg & Nakatani, 1996).

Since there is no standardized technique to estimate segmentation agreement, we developed our own approach based on percent agreement. We defined four types of segmentation agreement:

- *Perfect agreement (PA)*—Annotators completely agree on the segment boundaries.
- *Contiguous segments of the same type (ST)*—A segment of one annotator is divided into several segments of the same type by the other annotator. Segments are of the same type if they are marked with the same dialogue act tag and the

same addressee tag. An additional constraint is that segments are not labelled as parts of APs.

- *Unlabelled-DA (UDA)*—A segment of one annotator is divided into two segments by the other annotator where one of those segments is marked as *Unlabelled* and the other one with a dialogue act tag.
- *Conjunction-Floor(CF)*—Two adjacent segments differ only in a conjunction or a floor mechanism at the end of the first segment. The following example shows the segmentation agreement of this type:

(1) I can do that—but I need your help
(2) I can do that but—I need your help

The approach takes one annotator's segmentation as a reference (R) and compares it with the other annotator's segmentation (C) segment by segment. As a result, it gives a new segmentation (C') that represents the modification of (C) to match the reference segmentation (R) according to identified types of agreement. In addition to measuring segmentation agreement, the modified segmentation (C') is used for assessing reliability of dialogue act classification, addressee classification and adjacency pairs annotation. Table 3 shows overall segmentation results for each annotation group.

Most of the segmentation disagreements are of the following three types. First, while one annotator labelled a segment with the *Acknowledgement* tag, the other one included the segment in the dialogue act that follows. Second, while one annotator marked a segment with one of the response tags, the other annotator split the segment into a response and a statement that has a supportive function such as explanation, elaboration or clarification. Third, while one annotator split a segment into two or more segments labelled with the same dialogue act tag but different addressee tags, the other annotator marked it as one segment.

7.1.2 Reliability of dialogue act classification

Reliability of dialogue act classification is measured over those dialogue act segments for which both annotators agreed on their boundaries. Since the number of agreed segments for each R–C pair is different, we calculated reliability of dialogue act classification for each pair. The results are shown in Table 4. According to Krippendorff's scale annotators in each DA group reached an acceptable level of agreement that allows drawing tentative conclusions from the data.

We applied a single-category reliability test for each dialogue act tag to assess the extent to which one dialogue tag was confused with the other tags in the set. Table 5

Table 3 Segmentation agreement (R–C pair: reference annotator (R)–comparison annotator (C))

R–C	Agreement types						
	PA	ST	UDA	CFM	Agree	Total	Agree %
B–E	326	22	16	2	366	406	90.15
E–B	326	32	17	2	377	411	91.73
M–R	317	29	10	2	358	419	85.44
R–M	317	33	15	2	367	426	86.14

Table 4 Inter-annotator agreement on DA classification

Group	R–C pair	N	κ
M&R	M–R	358	0.70
	R–M	367	0.70
B&E	B–E	366	0.75
	E–B	377	0.77

Table 5 Single-category reliability for DA tags (Kappa values)

Category	B–E	M–R
Statement	0.82	0.72
Acknowledgement	0.87	0.75
Assessment/Appreciation	0.32	0.39
Information request	0.70	0.84
Open-ended question	0.74	0.84
Repetition request	1.00	1.00
Rhetorical question	0.00	0.66
Influencing-listeners-action	0.58	0.70
Committing-speaker-action	0.86	0.74
Positive response	0.70	0.52
Uncertain response	0.80	0.50
Negative response	0.67	0.61
Understanding check	0.32	–0.01
Other polite	0.00	–
Thanks	1.00	1.00
Follow me	–	– 0.003

shows the results of performing the Kappa tests for only one R–C pair in each DA group.

Annotators in the B&E group used different ranges of categories: the *Other polite* and *Rhetorical Question* categories, which occur rarely in the data, were employed only by annotator B. For that reason, Kappa values for these categories are zero. Negative Kappa values for *Understanding Check* and *Follow me* categories indicate that annotator agreement is below chance: in all cases where one annotator identifies one of these two categories, the other annotator does not. The results show an unacceptably low agreement on *Assessment/Appreciation* and *Understanding Check* categories in both groups. The *Assessment/Appreciation* category was mainly confused with *Positive Response* and *Statement* categories. The *Understanding Check* category was mostly confused with *Information Request* and *Statement* categories. Annotators in the M&R group reached a lower agreement on the responses tags than annotators in the B&E group. The responses tags were mostly confused with the *Statement* tag. Additionally, annotators in the M&R group had a little more difficulty distinguishing *Positive Response* from *Assessment/Appreciation* and *Acknowledgement*. The low Kappa value for the *Influencing-listener-actions* category in the B&R group is a result of the confusion with the *Statement* category.

7.2 Reliability of addressee annotation

As for dialogue act classification, reliability of addressee annotation is measured over those dialogue act segments for which both annotators agreed on their boundaries.

Table 6 Inter-annotator agreement on addressee annotation

Group	R–C pair	N	κ
M&R	M–R	358	0.68
	R–M	367	0.70
B&E	B–E	366	0.79
	E–B	377	0.81

The Kappa values for addressee annotation are shown in Table 6. The results show that annotators in the B&E group reached good agreement on addressee annotation, whereas annotators in the M&R group reached an acceptable level of agreement.

Annotators mainly disagreed on whether an individual or a group had been addressed. When annotators agreed that an individual had been addressed, they agreed in almost all cases which individual it had been. There were only a few instances in the data labelled with categories that represent subgroup addressing. In both DA groups, annotators failed to agree on those categories. Annotators had problems distinguishing subgroup addressing from addressing the group as a whole.

We measured single-category reliability for those addressee tags that represent individual and group addressing. Single-category reliability is measured using the Kappa test for one R–C pair in each group. Addressee values that consist of three participants such as p_0,p_1,p_3 or p_1,p_2,p_3 were grouped into one category that represents the whole audience (ALLP). Annotators in the B&E group reached a good agreement ($\kappa \geq 0.80$; $N = 369$) on all categories representing a single participant.

Agreement on ALLP was $\kappa = 0.77$. Annotators in the M&R group reached a lower agreement on each category than annotators in the B&E group. They had a little more difficulty distinguishing ALLP ($\kappa = 0.63$; $N = 366$) as well as p_3 ($\kappa = 0.59$; $N = 366$) from a remaining set of categories. For all other categories representing a single participant Kappa was $0.71 \leq \kappa < 0.80$.

7.3 Reliability of adjacency pairs annotation

According to our scheme for annotation of adjacency pairs, each dialogue act can be marked as a B-part of at most one and as an A-part of an arbitrary number of adjacency pairs. The sets of adjacency pairs produced by two annotators may differ in several ways. First, the annotators may disagree on dialogue acts that are marked as A-parts of adjacency pairs. Second, they may assign a different number of B-parts as well as different B-parts themselves to the same A-part.

Since there seems to be no standard associated metric for agreement on APs annotation in the literature, we developed a new approach that resembles a method for measuring reliability of co-reference annotation proposed in (Passonneau, 2004). The key of the approach is to represent annotated data as a form of categorical labelling in order to apply standard reliability metrics.

Adjacency pairs annotation can be seen as assigning to each dialogue act a *context* that represents the relations that the dialogue act has with surrounding dialogue acts. To encode the contexts of dialogue acts, we define a set of classes that contain related dialogue acts. For each A-part, all its B-parts are collected in one class. Therefore, a class is characterized with its A-part and a set of B-parts (b-set): $< a,b\text{set}(a) >$ where $b\text{set}(a) = \{b|(a,b) \in AP\}$. A dialogue act can belong to at most

two classes: a class containing the dialogue act as an A-part (A-class) and a class containing the dialogue act as a B-part (B-class). Thus, the complete context of a dialogue act is encoded with an AP label (L) that is compounded of its A-class and B-class (L=A-class|B-class).

Given a list of dialogue acts DA = [da_1,...,da_n], a class can be represented in two different ways: with fixed or relative position of the dialogue acts. The former encodes each dialog act in the class with the index of the dialog acts in the list. The latter encodes the dialogue acts in the class with relative positions to the dialogue act representing the A-part of the class. In this paper, we use the approach with relative positions because it significantly decreases the number of possible classes. In our encoding, each class of the labelled dialogue act da_i (A-class and B-class) has the form $< -n,O >$, where n is an offset of the labelled DA da_i from the A-part of the class and O is a set of offsets of the dialogue acts in the b-set from the A-part of the class. Note that for the A-class, n is always 0 since the labelled dialogue act is the A-part of the class. For the B-class, n is always positive because the labelled dialogue act is in the b-set and the A-part always precedes dialogue acts in the b-set. Thus, $-n$ refers to the dialogue act that is the A-part of the class. In the case where the labelled dialogue act is not an A-part or a B-part of an adjacency pair, one or both of the A-class and the B-class can be empty ($< 0,\{\} >$).

The proposed encoding makes patterns of disagreements between annotators directly visible. For example, (1) if one annotator marks the dialogue act 43 as an A-part of two adjacency pairs with B-parts 44 and 45, respectively, and the dialogue act 45 as an A-part of an adjacency pair with the B-part 47, and (2) the other annotator marks the dialogue act 44 as an A-part of an adjacency pair with the B-part 45 and the dialogue act 45 as an A-part of two adjacency pairs with B-parts 46 and 47, respectively, then the dialogue acts will be labelled as presented in Table 7. Fig. 2 illustrates the relation between the context of the dialogue act 45 and the AP label that encodes this context.

Encoding context in this way enables us to estimate *for each dialogue act* to what extent annotators agree on relating that dialogue act with surrounding dialogue acts in several ways: (1) as being an A-part related to a number of B-parts, (2) as being a B-part related to other B-parts with the same A-part and (3) not being related at all. It is to be noted that the context can be encoded in different ways as well. For example, it is possible to label each dialogue act that is marked as an A-part with its b-set. In this way, the actual disagreement is estimated only over A-parts. As context labels are not assigned to dialogue acts marked as B-parts, these dialogue acts would always be considered as agreed.

Agreement on APs annotation is measured over those dialogue acts for which annotators agreed on their boundaries. For computing agreement between annotators we use Krippendorff's α measure. This measure allows the usage of an

Table 7 An example of adjacency pairs annotation (C_1 and C_2: original AP annotations; $C_1(1)$ and $C_2(1)$: AP labels)

DA	C_1	C_2	$C_1(1)$	$C_2(1)$
43	1*a*2*a*		⟨0,{1,2}⟩|⟨0,{}⟩	⟨0,{}⟩|⟨0,{}⟩
44	1*b*	1*a*	⟨0,{}⟩|⟨ –1,{1,2}⟩	⟨0,{1}⟩|⟨0,{}⟩
45	3*a*2*b*	2*a*3*a*1*b*	⟨0,{2}⟩|⟨ –2,{1,2}⟩	⟨0,{1,2}⟩|⟨ –1,{1}⟩
46		2*b*	⟨0,{}⟩|⟨0,{}⟩	⟨0,{}⟩|⟨ –1,{1,2}⟩
47	3*b*	3*b*	⟨0,{}⟩|⟨ –2,{2}⟩	⟨0,{}⟩|⟨ –2,{1,2}⟩

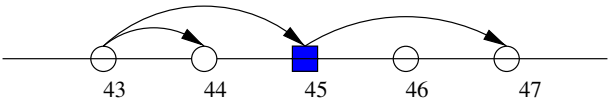

Fig. 2 A graphical representation of the context of dialogue act 45. The label that encodes this context is < 0,{2} > | < –2,{1,2} >

appropriate user defined distance metric on the AP labels. For nominal categories, the usual α distance metric (δ) is a binary function: $\delta = 0$ if categories are equal, otherwise $\delta = 1$. We need to use a more refined distance metric, one that is sensitive for partial agreement of annotators on the context they assign to a dialogue act. The agreement on the contexts is translated to agreements on the corresponding A-classes and B-classes. When annotators disagree, their disagreement should be penalized based on the difference between classes.

The intuition is that similarity of two classes with the same A-part depends not only on the number of elements in the intersection of their b-sets, but also on the size of both sets. Therefore, we define a distance metric δ' that uses the following similarity measure on sets:[6]

$$\mathrm{sim}(c_1, c_2) = \frac{2|c_1 \cap c_2|}{|c_1| + |c_2|} \tag{1}$$

The distance metric (δ') between the corresponding A-classes (or B-classes) of two AP labels is defined as:

$$\delta'(\langle -n_1, O_1\rangle, \langle -n_2, O_2\rangle) = 1, n_1 \neq n_2 \tag{2}$$

$$\delta'(\langle -n, O_1\rangle, \langle -n, O_2\rangle) = 1 - \mathrm{sim}(O_1, O_2) \tag{3}$$

The distance between two AP labels, $L_2 = A_1|B_1$ and $L_2 = A_2|B_2$, is defined as:

$$\delta_\lambda(L_1, L_2) = \lambda \cdot \delta'(A_1, A_2) + (1 - \lambda)\delta'(B_1, B_2), \tag{4}$$

where $\lambda \in [0,1]$ is a factor that determines the relative contribution of the distance between the corresponding classes the labels consist of.

Applying $\delta_{0.5}$ to the data of exactly one R–C pair in each group gave the following results: M–R: $\alpha = 0.71$ ($N = 260$), B–E: $\alpha = 0.83$ ($N = 322$). The most frequently occurring disagreement is when one annotator marks a dialogue act with the empty label, the other annotator with a non-empty one. If annotators agreed that a dialogue act is an A-part of an adjacency pair, they mostly agreed, either partially or fully, on the B-set of this dialogue act. In most cases, the confusion between (1) an AP label with both A-class and B-class non-empty and (2) an AP label with one of the classes empty is related to the disagreement on the DA tags assigned by annotators. This concerns the confusion between (i) *Statement* and *Assessment/Appreciation* tags, (ii) *Statement* and *Response* tags and (iii) *Understanding Check* and *Information Request* tags.

[6] The defined similarity measure is known as *Dice coefficient* (Manning & Schutze, 1999).

7.4 Reliability of gaze annotation

To evaluate reliability of gaze annotation, we first measured annotators agreement on marking the changes in gazed targets. Then, we measured agreement on labelling of time segments with gazed targets.

Marking the changes in gazed targets results in a segmentation of the time-line into non-overlapping, continuous segments that cover the whole input. In other words, the start time of a segment coincidences with the end time of the segment that precedes. A segment boundary indicates a change in gazed target.

The segmentation agreement is measured over all locations where any of the annotators marked a segment boundary. The number of locations where both annotators agree to some tolerance level is averaged over the total number of locations marked as a boundary. A tolerance level is introduced because the gaze annotation schema does not impose the requirement for a high precision on labelling changes in the gazed targets. It is defined to adjust the difference in whether a change is marked at the moment when the speaker starts changing the gaze direction or at the moment when the new target has been reached. It also adjusts the difference in the reaction of the annotators to the observed changes. Empirical analysis of the data shows that two points of the time-line can be considered equal with a tolerance level of 0.85 s.

The agreement on locations where any coder marked a segment boundary is 80.40% ($N = 939$). Annotators mostly disagreed on marking the cases when a participant briefly changes the gaze direction and then looks again at the previous target. Annotators reached very good agreement on gaze labelling ($\kappa = 0.95$) measured over those segments where boundaries were agreed.

8 Intra-annotator reliability

Intra-annotator reliability measures whether the results of a single annotator remain consistent over time. We assessed intra-annotator reliability of dialogue act and addressee annotation. One meeting from each data subset has been annotated twice by each annotator in the DA group over a period of three months. The results presented in Table 8 show that agreement on dialogue act annotation was good for each annotator indicating intra-annotator consistency in applying the dialogue act scheme. Furthermore, the results show that annotator R had a little more difficulty with addressee annotation than other annotators who reached good agreement.

Table 8 Intra-annotator agreement

Coder	Total	Agree	Segmentation	DA(κ)	ADD(κ)
E	110	104	94.54 %	0.83	0.88
B	107	104	97.20 %	0.89	0.81
M	73	64	87.67 %	0.81	0.87
R	77	72	93.51 %	0.85	0.76

9 Discussions and conclusions

We presented a multi-modal corpus of hand-annotated meeting dialogues that is designed for studying addressing behaviour in face-to-face conversations involving four participants. The corpus currently contains dialogue acts, addressees, adjacency pairs and gaze directions of meeting participants.

Annotators involved in the corpus design were able to reproduce the gaze labelling reliably. The annotations of dialogue acts and addressees were somewhat less reliable but still acceptable. Since there are only a few instances of subgroup addressing in the data and annotators failed to agree on them, the corpus cannot be used for exploring the patterns in addressing behaviour when a subgroup is addressed. In this paper, we have also presented a new approach for measuring reliability of adjacency pairs annotation. The key of the approach is to represent AP annotated data as a form of categorical labelling in order to apply standard reliability metrics.

The corpus has already been used for the development of models for automatic addressee prediction (Jovanovic, op den Akker, & Nijholt, 2006). Apart from addressing, the corpus can be exploited for studying other interesting aspects of conversations involving more than two participants. As the NXT stand-off XML format enables an easy extension of the corpus with new annotation layers without influencing existing annotations, the corpus can be extended to include, for example, coding of turn-taking mechanisms which would enable studying this aspect of conversational interaction independently as well as in relation to addressing.

Acknowledgements This work was partly supported by the European Union 6th FWP IST Integrated Project AMI (Augmented Multi-party Interaction, FP6-506811, publication AMI-160). We would like to thank Dennis Reidsma, Dennis Hofs, Lynn Packwood and the annotators who were involved in the corpus development. We are grateful to Klaus Krippendorff for useful discussions about reliability metrics.

References

Bakx, I., van Turnhout, K., & Terken, J. (2003). Facial orientation during multi-party interaction with information Kiosks. In *Proceedings of the 9th IFIP TC13 International Conference on Human–Computer Interaction (INTERACT)*, (pp. 701–704). Zurich, Switzerland

Burger, S., & Sloane, Z. (2004). The ISL meeting corpus: Categorical features of communicative group interactions. In *Proceedings of the NIST ICASSP Meeting Recognition Workshop*. Montreal, Canada

Carletta, J. (1996). Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics, 22*(2), 249–254

Carletta, J., Anderson, A. H., & Garrod, S. (2002). Seeing eye to eye: An account of grounding and understanding in work groups. *Cognitive Studies: Bulletin of the Japanese Cognitive Science Society, 9*(1), 1–20

Carletta, J., Ashby, S., Bourban, S. M., Flynn, M. G., Hain, T., Kadlec, J., Karaiskos, V., Kraaij, W., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska, A., McCowan, I., Post, W., Reidsma, D., & Wellner, P. (2006). The AMI meeting corpus: A pre-announcement. In S. Renals & S. Bengio (Eds.), *Machine learning for multimodal interaction, Second International Workshop, MLMI 2005, Edinburgh, UK, July, 2005, revised selected papers*, vol 3869 of *Lecture Notes in Computer Science* (pp. 28–39). Springer-Verlag. ISBN 3-540-32549-2

Carletta, J., Evert, S., Heid, U., Kilgour, J., Robertson, J., & Voormann, H. (2003). The NITE XML toolkit: Flexible annotation for multi-modal language data. *Behavior Research Methods, Instruments, and Computers, 35*(3), 353–363

Carletta, J., Isard, A., Isard, S., Kowtko, J., Doherty-Sneddon, G., & Anderson, A. (1997). The reliability of a dialogue structure coding scheme. *Computational Linguistics, 23*(1), 13–31

Chen, L., Rose, R. T., Parrill, F., Han, X., Tu, J., Huang, Z., Harper, M., Quek, F., McNeill, D., Tuttle, R., & Huang, T. (2006). VACE multimodal meeting corpus. In S. Renals & S. Bengio (Eds.), *Machine learning for multimodal interaction, Second International Workshop, MLMI 2005, Edinburgh, UK, July, 2005, revised selected papers*, vol 3869 of *Lecture Notes in Computer Science* (pp. 40–51). Springer-Verlag. ISBN 3-540-32549-2

Clark, H. H. & Carlson, T. B. (1992). Hearers and speech acts. In H. H. Clark (Ed.), *Arenas of Language Use* (pp. 205–247). University of Chicago Press and CSLI

Clark, A., & Popescu-Belis, A. (2004). Multi-level dialogue act tags. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue* (pp. 163–170). Cambridge, MA

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37–46

Dhillon, R., Bhagat, S., Carvey, H., & Shriberg, E. (2004). Meeting recorder project: Dialogue act labeling guide. Technical report TR-04-002, International Computer Science Institute (ICSI), Berkeley, CA, USA

Garofolo, J. S., Laprun, C. D., Michel, M., Stanford, V. M., & Tabassi, E. (2004) The NIST meeting room pilot corpus. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)* (pp. 1411–1414). Lisbon, Portugal

Goffman, E. (1981). Footing. In Goffman, E., (Ed.), *Forms of talk* (pp. 124–159). Philadelphia, PA: University of Pennsylvania Press

Goodwin, C. (1981). *Conversational organization: Interaction between speakers and hearers*. New York: Academic Press

Hirschberg, J., & Nakatani, C. (1996). A prosodic analysis of discourse segments in direction-giving monologues. In *In Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL)* (pp 286–293). Santa Cruz, California

Janin, A., Ang, J., Bhagat, S., Dhillon, R., Edwards, J., Macias-Guarasa, J., Morgan, N., Peskin, B., Shriberg, E., Stolcke, A., Wooters, C., & Wrede, B. (2004). The ICSI meeting project: Resources and research. In *Proceedings of the NIST ICASSP Meeting Recognition Workshop*. Montreal, Canada

Jovanovic, N., & op den Akker, R. (2004). Towards automatic addressee identification in multi-party dialogues. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue* (pp. 89–92). Cambridge, MA

Jovanovic, N., op den Akker, R., & Nijholt, A. (2006). Addressee identification in face-to-face meetings. In *Proceeding of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)* (pp. 169–176). Trento, Italy

Jurafsky, D., Shriberg, L., & Biasca, D. (1997). Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation Coders Manual, Draft 13. Technical report 97–02, University of Colorado, The Institute of Cognitive Science, Boulder, CO

Katzenmaier, M., Stiefelhagen, R., & Schultz, T. (2004). Identifying the addressee in human–human–robot interactions based on head pose and speech. In *Proceedings of the International Conference on Multimodal Interfaces (ICMI)* (pp. 144–151). State College, PA

Krippendorff, K. (1980). *Content analysis: An introduction to its methodology*. Sage Publications, Beverly Hills, CA

Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research, 30*(3), 411–433

Manning, C., & Schutze, H. (1999). *Foundations of statistical natural language processing*. The MIT Press, Cambridge, MA

Moore, D. (2002). The IDIAP smart meeting room. Technical report IDIAP-COM-07, IDIAP, Martigny, Switzerland

Otsuka, K., Takemae, Y., Yamato, J., & Murase, H. (2005). A probabilistic inference of multiparty-conversation structure based on Markov-switching models of gaze patterns, head directions, and utterances. In *Proceedings of the International Conference on Multimodal Interfaces (ICMI)* (pp. 191–198). Trento, Italy

Passonneau, R. (2004). Computing reliability for coreference annotation. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)* (pp. 1503–1506). Lisbon, Portugal

Passonneau, R. & Litman, D. (1997). Discourse segmentation by human and automated means. *Computational Linguistics, 23*(1), 103–139

Reidsma, D., Hofs, D., & Jovanovic, N. (2005). A presentation of a set of new annotation tools based on the NXT API. In *Proceedings of Measuring Behavior* (pp. 512–513). Wageningen, The Netherlands

Shriberg, E., Dhillon, R., Bhagat, S., Ang, J., & Carvey, H. (2004). The ICSI meeting recorder dialog act (MRDA) corpus. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue* (pp. 97–100). Boston, USA

Traum, D. (2004). Issues in multi-party dialogues. In F. Dignum (Ed.), *Advances in agent communication* (pp. 201–211). Springer-Verlag LNCS

van Turnhout, K., Terken, J., Bakx, I., & Eggen, B. (2005). Identifying the intended addressee in mixed human–human and human–computer interaction from non-verbal features. In *Proceedings of the International Conference on Multimodal Interfaces (ICMI)* (pp.175–182). Trento, Italy