A corpus for studying addressing behavior in multi-party dialogues

Natasa Jovanovic
University of Twente
PO Box 217 Enschede
the Netherlands
natasa@ewi.utwente.nl

Rieks op den Akker
University of Twente
PO Box 217 Enschede
the Netherlands
infrieks@ewi.utwente.nl

Anton Nijholt
University of Twente
PO Box 217 Enschede
the Netherlands
A.Nijholt@ewi.utwente.nl

Abstract

This paper describes a multi-modal corpus of hand-annotated meeting dialogues that was designed for studying addressing behavior 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 annotation schema, annotation tools and annotation process itself. Then, we analyze the reproducibility and stability of the annotation schema.

Keywords: multimodal resources, coding schemes, reliability measures

1 Introduction

Current tendencies in modeling 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 (Goffman, 1981; Clark and Carlson, 1992; Traum, 2003).

Addressing is carried out through various communication channels, e.g. speech, gesture or gaze. Existing corpora, such as ICSI and ISL meeting corpora-currently widely used to study linguistic phenomena in natural meetings (Janin et al., 2004; Burger and Sloane, 2004), are limited to audio data only. To explore interaction patterns in addressing behavior and to develop statistical models for automatic addressee prediction, we need a corpus that contains video data as well.

In this paper, we describe a new multi-modal corpus of hand-annotated meeting dialogues, designed for studying addressing behavior in small 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 media file server[3]. Currently, the corpus contains hand-annotated dialogue acts, adjacency pairs, addressees and gaze directions of meeting participants.

This paper reports the reliability of the overall annotation schema as well as a detailed analysis of detected sources of unreliability.

The overall annotation schema is presented in Section 3. Annotation tools used in the creation of the corpus and the annotation process itself are described in Section 4 and Section 5 respectively. In Section 6 we give an overview of the applied reliability tests and measures as well as methods for detecting sources of unreliability. The results of these tests and methods applied on dialogue act annotation, addressee annotation, adjacency pairs annotation and gaze annotation are given in Section 7 and Section 8.

[1] http://www.m4project.org
[2] http://www.amiproject.org
[3] http://mmm.idiap.ch

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 and video recording devices. Of the 12 meetings, 10 were recorded within the scope of the M4 project. The meetings are scripted as to which actions the participants will undertake, but not as to what they will say. Although the meetings are inappropriate for research into richer meeting analysis due to their constrained nature, they allow us to examine observable patterns of addressing behavior in small group discussions. More natural, scenario-based, meetings are being recording at IDIAP in the scope of the AMI project. One of the new pilot meetings 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 for the exploration of argumentative structures in meeting dialogues.

There are 23 participants in the corpus. Each meeting consists of 4 participants. The total amount of recorded data is approximately 75 minutes.

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 schema is based on the model presented in (Jovanovic and 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 behavior 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, currently, the scheme contains only annotations of dialogue acts, adjacency pairs, addressees and gaze direction.

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, are segmented in advance using prosody, pause and syntactical information. In our schema, 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 et al., 2004). 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 and 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.

In our dialogue act annotation scheme each utterance is marked as *Unlabeled* 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 as well as the mapping between our tag set and the MRDA set is shown in Table 1.

DA tag set	MRDA
Statements	
s Statement	s Statement
Questions	
q Information-Request	Wh-question, Y/N question, OR-question, Or Clause After Y/N question
qo Open-ended Question	Open-ended questions
qh Rhetorical Question	Rhetorical Questions
Backchannels and Ack.	
bk Acknowledgement	Acknowledgment,Backchannel
ba Assessment/Appreciation	Assessment/Appreciation
Responses	
rp Positive response	(Partial)Accept, Affirmative Answer
rn Negative response	(Partial)Reject, Dispreferred and Negative Answer
ru Uncertain response	Maybe , No Knowledge
Action Motivators	
al Influencing-listeners-action	Command, Suggestion
as Committing-speaker-action	Commitment, *Suggestion*
Checks	
f "Follow Me"	"Follow Me"
br Repetition Request	Repetition Request
bu Understanding Check	Understanding Check
Politeness Mechanisms	
fa Apology	Apology
ft Thanks	Thanks
fo Other polite	Downplayer,Sympathy, Welcome

Table 1: Dialogue act tag set

3.2 Adjacency pairs

Adjacency pairs (APs) are 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.

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 behavior 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
- $Unknown$

where $x, y, z \in \{0, 1, 2, 3\}$; 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.

3.4 Gaze direction

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

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 Labeling) tools (Reidsma et al., 2005). The DACoder supports annotation of dialogue acts, addressees and any kind of relations between dialogue acts. The CSL tool supports labeling of time-aligned annotation layers directly related to the signal files. Any annotation layer that consists of simple labeling of non-overlapping segments of the time line can be coded using this tool (e.g. gaze directions, postures and emotions).

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.

5 Annotation procedure

Six trained annotators were involved in the corpus creation. They were divided into two groups: the DA (Dialogue Act) group and the VL (Video Labeling) 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 6 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.

The annotators performed their tasks following different procedures. Two annotators from the DA group annotated dialogue acts, addressee and adjacency pairs separately, whereas the others annotated dialogue acts and addressees in one pass and adjacency pairs in the other pass. One annotator from the VL group annotated gaze direction in real-time, while the other annotator annotated gaze direction off-line. For the DA group, labeling time of 5 minutes of meeting data averaged about two and a half hours. Real-time labeling of gaze direction for four meeting participants averaged about 20 minutes for 5 minutes of meeting data, whereas off-line annotation averaged about 5 hours for the same amount of data.

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

6.1 Kappa vs. Alpha

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 (Cohen, 1960; Carletta, 1996). In recent years, there have been some discussions about the usage of Kappa as an appropriate reliability metric. The main problem when employing Kappa is that it actually depends on marginal distributions. As shown in (Krippendorff, 2004), Kappa expected disagreement is a function of the individual coder preferences for the categories, and not of the proportions of categories in the data.

An agreement coefficient that does not have this inadequacy is Krippendorff's Alpha (α) (Krippendorff, 1980). Since Alpha measures properties of the data and not coders' preferences, it is easily interpretable compared to Kappa. When a sample size is large and coders agree on their use of categories, $\kappa = \alpha$ (Krippendorff, 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 labeling of annotation units with categories. Therefore, we developed our own approach that represents annotated APs in a form of categorical labeling and measures agreement on APs annotation using Alpha.

For the evaluation of Alpha and Kappa values, we used the 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.

6.2 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 use 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 inter-annotator reliability of the annotation schema applied on the M4 meeting data.

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.

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 et al., 2004), precision and recall (Passonneau and Litman, 1997), and κ (Carletta et al., 1997; Hirschberg and 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 labeled as parts of APs.
- Unlabeled-DA (UDA)-A segment of one annotator is divided into two segments by the other annotator where one of those segments is marked as *Unlabled* 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 2 shows overall segmentation results for each annotation group.

	Agreement types						
R-C	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 2: Segmentation agreement (R-C pair: Reference annotator (R)-Comparison annotator)

Most of the segmentation disagreements are of the following three types. First, while one annotator labeled a segment with the *Acknowledgment* 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 labeled 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 3. According to Krippendorff's scale annotators in each DA group reached an acceptable level of agreement that allows drawing tentative conclusions from data.

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 3: Inter-annotator agreement on DA classification

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

Category	B-E	M-R
Statement	0.82	0.72
Acknowledgment	0.87	0.75
Assessment/Appreiciation	0.32	0.39
Information-Request	0.70	0.84
Open-ended question	0.74	0.84
Repetition request	1.00	1.00
Rhetorical questions	0.00	0.66
Influencing-listeners-actions	0.58	0.70
Committing-speaker-actions	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

Table 4: Single-category reliability for DA tags (Kappa values)

Annotators in the B&E group used different ranges of categories. For that reason, Kappa values of the categories that are used by only one annotator are zero. Negative Kappa values for *Understanding check* and *Follow-me* categories indicate that annotator agreement is below the 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 merely 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.

The Kappa values for addressee annotation are shown in Table 5. 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 that allows drawing tentative conclusions from data.

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

Table 5: Inter-annotator agreement on addressee annotation

We measured single-category reliability 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 $0.71 \leq \kappa < 0.80$. There were only a few instances in the data labeled with categories that represent a 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.

7.3 Reliability of adjacency pairs annotation

According to our schema 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. Secondly, 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 labeling 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): $\langle a, bset(a) \rangle$ where $bset(a) = \{b | (a, b) \in AP\}$. 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 that is compounded of its A-class and B-class ($L = A - class | B - class$).

Given a list of dialogue acts $DA = [da_1, \ldots, da_n]$, 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

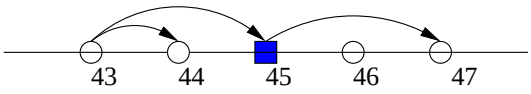

Figure 1: A graphical representation of the context of dialogue act 45. The label that encodes this context is $< 0, \{2\} > | < -2, \{1, 2\} >$

DA	C_1	C_2	$C_1(1)$	$C_2(1)$
43	$1a2a$		$\langle 0, \{1, 2\}\rangle \vert \langle 0, \{\}\rangle$	$\langle 0, \{\}\rangle \vert \langle 0, \{\}\rangle$
44	$1b$	$1a$	$\langle 0, \{\}\rangle \vert \langle -1, \{1, 2\}\rangle$	$\langle 0, \{1\}\rangle \vert \langle 0, \{\}\rangle$
45	$3a2b$	$2a3a1b$	$\langle 0, \{2\}\rangle \vert \langle -2, \{1, 2\}\rangle$	$\langle 0, \{1, 2\}\rangle \vert \langle -1, \{1\}\rangle$
46		$2b$	$\langle 0, \{\}\rangle \vert \langle 0, \{\}\rangle$	$\langle 0, \{\}\rangle \vert \langle -1, \{1, 2\}\rangle$
47	$3b$	$3b$	$\langle 0, \{\}\rangle \vert \langle -2, \{2\}\rangle$	$\langle 0, \{\}\rangle \vert \langle -2, \{1, 2\}\rangle$

Table 6: An example of adjacency pairs annotation (C_1 and C_2: original AP annotations; $C_1(1)$ and $C_2(1)$: AP labels)

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 labeled dialogue act da_i (A-class and B-class) has the form $\langle -n, O\rangle$, where n is an offset of the labeled 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 labeled dialogue act is the A-part of the class. For the B-class, n is always a positive because the labeled 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 labeled 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 ($\langle 0, \{\}\rangle$).

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 labeled as presented in Table 6. Figure 1 illustrates the relation between the context of the dialogue act 45 and the AP label that encodes this context.

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 appropriate user defined distance metric on the AP labels. For nominal categories, the usual α distance metric (δ) is a binary function: $\delta = 1$ if categories are equal, otherwise $\delta = 0$. 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 on the number of elements in the intersection as well as on the number of elements in the union of their b-sets. Therefore, we define a distance metric δ' that uses the following similarity measure on sets[4]:

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

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

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

$$\delta'(\langle -n, O_1\rangle, \langle -n, O_2\rangle) = 1 - sim(O_1, O_2)$$

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

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

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:

[4]Known as *Dice coefficient*, see (Manning and Schutze, 1999)

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 (iii) Understanding check and Information Request tags.

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 labeling 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 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 7 show that agreement on dialogue act annotation was good for each annotator indicating intra-annotator consistency in applying the dialogue act schema. Furthermore, the results show that annotator R had a little more difficulty with addressee annotation than other annotators that reached good 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

Table 7: Intra-annotator agreement

9 Conclusion

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

Annotators involved in the corpus design were able to reproduce the gaze labeling reliably. The annotations of dialogue acts and addresses were somewhat less reliable but still acceptable. Since there are only 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 behavior 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.

Apart from addressing, the corpus can be exploited for studying other interesting aspects of conversations involving more than two participants. The NXT stand-off XML format enables an easy extension of the corpus with new annotation layers of different modalities.

Acknowledgments

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

References

S. Burger and Z. Sloane. 2004. The isl meeting corpus: Categorical features of communicative group interactions. In *Proc. of ICASSP 2004 Meeting Recognition Workshop*.

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

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

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

H.H. Clark and T.B. Carlson. 1992. Hearers and speech acts. *Arenas of Language Use (H.H.Clark ed.)*.

Alexander Clark and Andrei Popescu-Belis. 2004. Multi-level dialogue act tags. In *Proc. of 5th SIGdial Workshop on Discourse and Dialogue*.

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

R. Dhillon, S. Bhagat, H. Carvey, and E. Shriberg. 2004. Meeting recorder project: Dialogue act labeling guide. Technical Report TR-04-002, ICSI Speech Group, Berkeley, USA.

Erving Goffman. 1981. Footing. In *Forms of Talk*, pages 124–159. University of Pennsylvania Press.

J. Hirschberg and C.H. Nakatani. 1996. A prosodic analysis of discourse segments in direction-giving monologues. In *In Proc. of the 34th Annual Meeting of the Association for Computational Linguistics*.

A. Janin, J. Ang, S. Bhagat, R. Dhillon, J. Edwards, J. Macias-Guarasa, N. Morgan, B. Peskin, E. Shriberg, A. Stolcke, C. Wooters, and B. Wrede. 2004. The icsi meeting project: Resources and research. In *Proc. of ICASSP 2004 Meeting Recognition Workshop*.

N. Jovanovic and R. op den Akker. 2004. Towards automatic addressee identification in multi-party dialogues. In *Proc. of 5th SIGdial Workshop on Discourse and Dialogue*.

K. Krippendorff. 1980. *Content analysis: An introduction to its methodology*. Sage Publications.

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

C.D. Manning and H. Schutze. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press.

D. Moore. 2002. The idiap smart meeting room. Technical Report IDIAP-COM-07, IDIAP.

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

R.J. Passonneau. 2004. Computing reliability for coreference annotation. In *Proc. of LREC*.

D. Reidsma, D.H.W. Hofs, and N. Jovanovic. 2005. A presentation of a set of new annotation tools based on the nxt api. In *In Proc. Measuring Behavior*.

E. Shriberg, R. Dhillon, S. Bhagat, J.Ang, and H. Carvey. 2004. The icsi meeting recorder dialog act (mrda) corpus. In *Proc. of 5th SIGdial Workshop on Discourse and Dialogue*.

David Traum. 2003. Issues in multi-party dialogues. *Advances in Agent Communication In (F. Dignum, ed.)*.