TO WHOM IT MAY CONCERN

ADDRESSEE IDENTIFICATION IN FACE-TO-FACE MEETINGS

DISSERTATION

to obtain
the doctor's degree at the University of Twente,
on the authority of the rector magnificus,
prof. dr. W. H. M. Zijm,
on account of the decision of the graduation committee
to be publicly defended
on Wednesday, March 14, 2007 at 15:00

by

Nataša Jovanović

born on October 11, 1976
in Užice, Serbia

This dissertation is approved by:

Prof. dr. ir. Anton Nijholt (promotor)
Dr. ir. H.J.A. op den Akker (assistant-promotor)

Summary

This thesis is concerned with automatic addressee identification in face-to-face meetings.

The first part of the thesis is devoted to gaining theoretical insights into addressing based on the outcomes of the research in conversational and interaction analysis. The multimodal nature of addressing as well as its context sensitivity poses challenges not only for computational systems but also for humans in determining who is being addressed by the speaker. This thesis addresses both challenges.

The second part of the thesis describes two meeting corpora employed in our study. The corpora were developed using annotation schemes designed in the interaction between data observed and theoretical insights into addressing obtained from the literature. To assess the credibility of the annotated data, the thesis provides an exhaustive reliability analysis of the annotation schemes. A detailed investigation of the problems human observers had in determining who is being addressed by the speaker shows that it is intrinsically difficult to distinguish between group and individual addressing.

The third part of the thesis deals with the development of a computational model for automatic addressee identification using Bayesian Networks. Features employed to model addressing are obtained from speech and gaze communication channels as well as from conversational and meeting contexts. Conversational context features are shown to be the most valuable, whereas utterance features are found to be the least reliable cues for addressee prediction. Evaluation of several types of Bayesian Network classifiers indicates that Bayesian Networks are effective computational models for the task of addressee prediction. The highest accuracies of addressee classifiers (about 80%) are achieved by combining features from all available resources. The features used to model addressing are in a sense ideal because they are obtained from human-annotated data. As the first step in the automation of the addressee detection process, the thesis proposes a Dynamic Bayesian Network model. The evaluation of performances of addressee classifiers relying on fully automatic features remains one of the tasks for future research.

To Whom It May Concern

Addressee Identification in Face-to-Face Meetings

Nataša Jovanović

PhD dissertation committee:

Chairman:
Prof. dr. ir. A.J. Mouthaan

Promotor:
Prof. dr. ir. A. Nijholt

Assistant-promotor:
Dr. ir. H.J.A. op den Akker

Members:
Prof. dr. T. Becker, DFKI GmbH, Germany
Prof. dr. M. Pantic, Universiteit Twente, The Nederlands
Prof. dr. F.M.G de Jong, Universiteit Twente, The Nederlands
Prof. dr. G.C van der Veer, Vrije Universiteit Amsterdam, The Nederlands
Dr. J.M.B Terken, Technische Universiteit Eindhoven, The Nederlands
Prof. dr. D. Traum, USC Institute for Creative Technology, USA

CTIT Ph.D. thesis Series No. 07-96
Center for Telematics and Information Technology (CTIT)
P.O. Box 217 - 7500 AE Enschede - The Netherlands

SIKS Dissertation Series No. 2007-07
The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Graduate School for Information and Knowledge Systems.

ISBN: 978-90-365-2470-4
ISSN: 1381-3617

Cover design: Ronald Poppe & Nataša Jovanović
Printed by: Wöhrmann Print Service, Zutphen, The Netherlands

TO WHOM IT MAY CONCERN

ADDRESSEE IDENTIFICATION IN FACE-TO-FACE MEETINGS

DISSERTATION

to obtain
the doctor's degree at the University of Twente,
on the authority of the rector magnificus,
prof. dr. W. H. M. Zijm,
on account of the decision of the graduation committee
to be publicly defended
on Wednesday, March 14, 2007 at 15:00

by

Nataša Jovanović

born on October 11, 1976
in Užice, Serbia

This dissertation is approved by:

Prof. dr. ir. Anton Nijholt (promotor)
Dr. ir. H.J.A. op den Akker (assistant-promotor)

to my parents and my brother

Contents

Chapter 1

Introduction

1.1 Meetings and meetings support

Meetings constitute an important part of modern organizations. People meet to exchange ideas, share information, negotiate alternatives, collaboratively solve problems or make decisions. Meetings as such are embedded in a larger work process. In many cases, a work process is carried out through a series of meetings involving not always the same group of participants. Furthermore, due to time and space constraints, people are not always able to attend all of the meetings they are supposed to.

Although essential and increasing in numbers, meetings are, as many studies reveal, neither as efficient nor as productive as they might be. An overview of these studies is given in (Romano and Nunamaker, 2001). Inefficiency is manifested mostly through process loss and information loss. The former is caused by, for example, ineffective leadership, lack of participation or irrelevance of topics discussed (Romano and Nunamaker, 2001). The latter is, however, a result of a failure to record important information, insights, decisions and actions assigned to participants present as well as to non-attendees (Moran et al., 1997; Whittaker et al., 2006). The loss of information may affect future individual and collaborative actions, which in turn may have a negative impact on the outcome of the work process itself.

Traditionally meetings are documented by taking minutes which contain a minimal description of group decisions and collectively agreed actions during the meeting (Whittaker et al., 2006). Analyzing meeting recording practices, Whittaker et al. (2006) highlighted several drawbacks of minutes among which the most prominent are unavailability, inaccuracy and the lack of the informational context for the participants to carry out their own actions. Meeting participants usually respond to these limitations by taking their own notes. One of the most significant limitations of personal notes, as observed by the same authors, is that taking notes reduces the ability of the note-taker to contribute to discussions. Since they are usually cryptic and incomprehensive, non-attendees have problems in understanding other's personal notes. The user survey presented in (Banerjee and Rudnicky, 2006) confirms these findings. Additionally, the survey has shown that, when available documents were not sufficient to obtain the information of interest, participants usually asked someone who attended the meeting. However, this was not a guarantee that information received was accurate since it had been shown that participants may have different recollections of what went on in a meeting. Moreover, non-attendees' understanding of the meetings they missed was far from perfect even after considering all of these information resources.

All these findings obviously indicate that there is a need for the development of technology to support meetings regarding both the content and process. At this point, it is worth mentioning that a lot of work has already been done in facilitating meetings by transcending space, that is, by enabling non-collocated participants to conduct meetings. The technology has been invented to allow different modes of mediated communication such as audio, video or web conferencing. Here we focus on conventional meetings, their limitations and the technology needed to overcome those limitations.

With the advances in multimedia technologies, it has become feasible to capture all aspects of multimodal multi-party interaction taking place in meetings as well as the supporting meeting artifacts such as notes or slides. Multimedia meeting capturing obviously overcomes most of the limitations of the traditional recording practices mentioned above. To take benefit from meeting recordings two types of technology are needed: technology for automatic analysis of meeting content and technology for adequate access to meeting data (e.g. browsing, querying, filtering). Automatic meeting analysis is a rapidly emerging field, as the number of recent research projects seems to confirm. It comprises development of a wide range of technologies focusing on different aspects of interaction among meeting participants such as speech recognition, participant and speaker identification, topic detection and structuring, discussion modeling, detection of argumentative structures, summarization, identification of meeting events, emotion detection and so on. An overview of the relevant technologies concerning automatic meeting analysis with the reference to relatively recent literature is given in (Gatica-Perez et al., 2005).

Systems specifically designed to enable users to efficiently view and access automatically analyzed content of previously recorded meetings are known as meeting browsers. A meeting browser may, for example, support users in retrieving relevant information from the previous meetings as a preparation for a new meeting. It may also assist users during the meeting in accessing, for example, information on the same topic from previous meetings or from a previous part on the same meeting. A recent review of the existing meeting browsers is given in (Tucker and Whittaker, 2005). To provide useful support to meetings, browsers must satisfy the needs of potential users. A number of user studies were performed to explore user requirements for the exploitation of the meeting data (Lisowska, 2003; Jaimes et al., 2004; Banerjee et al., 2005; Cremers et al., 2005; Pallotta et al., 2006). The following queries were found, among other things, to be of interests for the potential users of meetings browsers:

- *To whom are action items assigned and by when?*
- *What tasks were assigned to me?*

- *Was there any discussion between participants X and Y?*

In order to answer these queries, some sort of understanding of conversational structure is required. Furthermore, different types of information need to be extracted from the meeting content in order to provide a complete answer to each of them. However, what is common for all these queries is that in order to be answered they need information about **who is being addressed** at the particular instance. Obviously, automatically generated meeting minutes, if available, should provide the required information concerning the action items. In that case, addressee information is substantial for automatic generation of action items components of meeting minutes (Purver et al., 2006).

Up to now, automatic understanding of meeting conversation was mostly focused on identifying who speaks, what is said, what type of communicative act the speaker performs with his utterance. However, the question to whom the speaker addresses this act remained unanswered. In this thesis, we address this issue by developing a model that provides an answer to this question.

There are many applications related to meeting research that could benefit from studying addressing in human-human interactions. So far, we have only discussed meeting browsers as tools to support meetings. However, participants in the meetings, regardless of whether they are in a joint physical space or participate remotely, can be supported by the environment itself and this support can be realized by introducing virtual agents into the environment to act as meeting assistants (Nijholt et al., 2006; Rienks et al., 2007; Purver et al., 2005). They can assist participants both personally and as a group in a variety of tasks such as providing relevant information for completing tasks, making a record of meetings or guarding the agenda. Meeting assistants may not be active participants in the discussion but, to assist actual participants in meetings, they need to *understand* natural interaction taking place among the participants. Also, when modeled as active participants in conversation, these agents need to *recognize* when they are being addressed and how they should *address* people in the environment.

Although the research presented in this thesis has been conducted in the context of smart meeting rooms, the knowledge gained by studying addressing among participants in meetings can be beneficial for modeling interaction in various other sensor-equipped environments. In ambient intelligence research the assumption is that computing intelligence will be embedded in our daily home, office and recreational environments and that the environments and their devices will offer social and intelligent interfaces. The interfaces, whether they are using built-in cameras and microphones or other kinds of sensors, perceive events and activities. The interpretation of what is perceived allows them to be reactive and pro-active, to predict and to anticipate. People meet in these environments, they interact with each other and with smart devices (e.g. humanoid robots) that are also present. Knowledge about addressing in multi-party interaction in such environments is needed for both the environment and its devices to be able to interpret the interactions and to be able to support human activities in these environments (Nijholt, 2007).

1.2 Addressing - bridging a gap between two-party and multi-party interaction modeling

Addressing is an aspect of every form of communication. It represents a form of orientation and directionality of the act the current actor performs toward the particular others who are involved in interaction. In conversational communication involving two participants, the hearer is almost always the addressee of the speech act that the speaker performs; the speaker may also talk to himself. However, in multi-party conversation, addressing becomes a real issue. In multi-party interaction, participants may, for example, be addressed (*addressee*) or unaddressed (*side-participants*) or they may overhear the conversation without taking part in it (*bystanders*).

The notion of addressee as well as mechanisms that speakers use in order to make clear to whom they are addressing their speech has been extensively studied by conversational analysts and social psychologists (Goffman, 1981a; Goodwin, 1981; Clark and Carlson, 1992). Moreover, these phenomena were investigated as a part of a more general concept that characterizes conversational interaction involving more than two participants - the *concept of participation* (Goffman, 1981a). Focusing their study on multi-party interaction, researchers in these fields emphasized that conversation is not only organized as a sequence of actions participants perform in speaking but also across different parties present. Research on participation has been mostly focused on (1) decomposition of traditional notions of speaker and hearer into a set of participation roles as well as on (2) specification of the means that speakers use in order to designate which hearers are to take which roles. The latter is refer to as *the recipient design principle* (Sacks et al., 1974; Clark and Carlson, 1992). A form of recipient design that is concerned with mechanisms that speakers use in order to designate which participants are to take the addressee role is known as *addressee design* (Clark and Carlson, 1992).

There is agreement among the researchers in these fields that the notion of addressee as well as the addressee design are complicated, significant and not very well explored. The complexity of addressing is influenced, in the first place, by its *multimodal* nature: addressing is carried out through various communication channels such as speech, gaze, hand and head gestures. Furthermore, in most cases addressing is not performed explicitly but tacitly. It depends on the course of the conversation, the status of attention of interacting participants, their involvement in discussion as well as on what the participants know about each other's state of the knowledge whether explicit addressing is asked for. This is what we refer to as *context sensitivity* of the addressing practices. Other elements of context that have an effect on addressee design are the participants' physical arrangement and the presence of various attention 'distracters' in the conversational environment. All these clearly impose a challenge for automatic modelling of addressing in multi-party interaction.

Up to recent years, the research fields concerned with computational modeling of conversational interaction in various contexts were focused exclusively on the two-party model. Therefore, addressing was not considered as a relevant issue. However, current changes in research tradition have shown tendencies to move from a two-party model to a multi-party model which resulted in the first attempts to model addressing. The research on addressee identification has been conducted in the context of mixed human-human and human-computer interaction (Bakx et al., 2003; van Turnhout et al., 2005; Lunsford and Oviatt, 2006), human-human robot interaction (Katzenmaier et al., 2004), mixed human-agents and multi-agent interaction (Traum, 2004) and multi-party conversation (Otsuka et al., 2005).

As previously stated, the work presented in this thesis is focused on modeling addressing in face-to-face meetings. Meetings as a special context of talk have their own properties that differ from ordinary conversations. For example, during a meeting participants may be engaged in different types of activities such as giving a presentation, writing notes or closing the meeting. This adds an additional complexity to modelling addressing, since the same kind of addressing behavior may be interpreted differently for different meeting activities. For example, gazing at a participant in the audience during a presentation usually does not mean that the participant being gazed at is addressed whereas the same kind of behavior during a discussion can be a strong indicator that the gazed participant is the addressed one. This example is illustrated in Figure 1.1

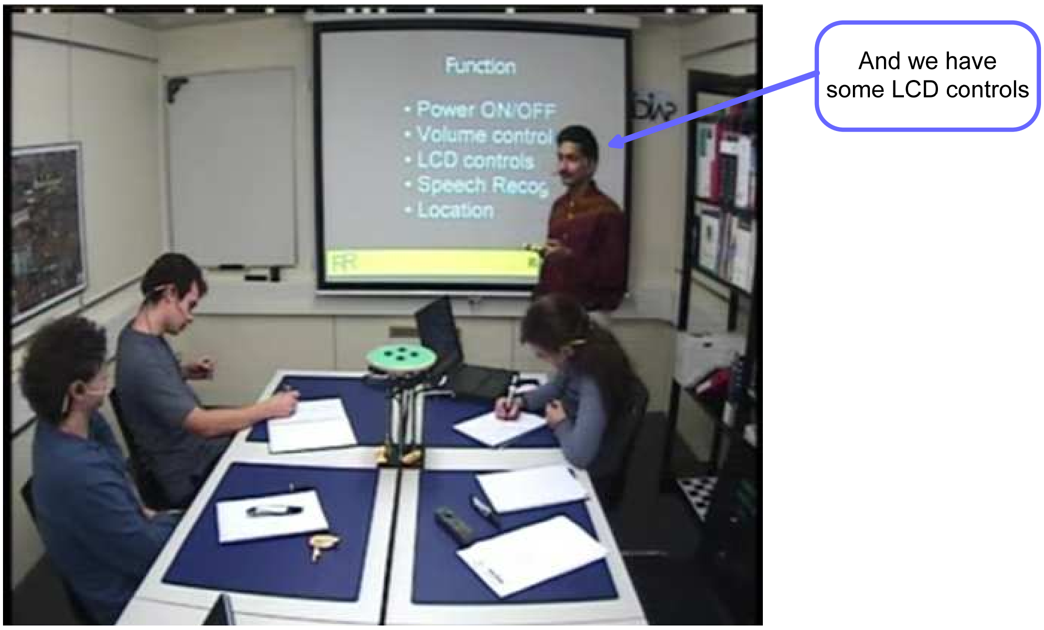


Figure 1.1: An example of addressing in meetings - the presenter is addressing the group

Research on small group discussions presented in (Carletta et al., 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. Due to their interactive nature, small group discussions are more appropriate for exploring interaction patterns of addressing behavior. Therefore, we focus our research on small group meetings. Moreover, we restrict ourselves to four-participants meetings which has been motivated by the choice of meeting collections we employed in our study, as will be justified in the following section.

1.3 Research goals and research approach

The research presented in this thesis has been conducted with two main goals. The first goal is to gain knowledge in how addressing is accomplished in face-to-face meetings. In order to achieve this goal several questions need to be answered:

- how is the addressee role distinguished from other participation roles in conversation?
- which addressing mechanisms do speakers employ in order to make clear to whom they are addressing their speech?
- how do addressing mechanisms depend on conversational context and meeting context?
- how is addressing related to other practices operating in conversation, namely, to turn-taking design and adjacency pairs organization?

To address these questions we survey the relevant literature regarding (1) sequential organization of conversational interaction (e.g. turn-taking design, adjacency pair organization and overall organization) and (2) organization of conversation across participants.

We further aim to develop a computational model for addressee identification in face-to-face meetings. The purpose of such a model is to identify for each communicative act that a speaker performs in speaking who is being addressed by that speaker based on a set of features extracted from the identified addressing mechanisms and context. This is known as a classification problem. As the addressing process in a particular situation is represented with a set of extracted features, modeling this process suffers from the loss of information and thus for uncertainty. To deal with uncertainty, we have adopted a probabilistic approach. More specifically, we have chosen Bayesian Networks as a probabilistic framework. Regarding the development of a computational model for addressee identification, two main lines of research can be distinguished:

- finding relevant features for addressee identification in face-to-face meetings
- finding the appropriate Bayesian Network models for this type of task.

The features employed for addressee identification are obtained from several resources: speech, gaze, conversational and meeting context.

To develop models for automatic addressee identification in meetings, we need a collection of audio and video meeting recordings annotated with addressee information as well as with a set of phenomena relevant to addressing (e.g. information who is looking at whom during the meeting). In this study, we make use of two meeting corpora developed in the course of two European projects in which contexts the research presented in this thesis has been conducted: the M4 (MultiModal Meeting Manager)[1] and AMI (Augmented Multiparty Interaction)[2] projects. Both projects were concerned with the development of technology to support meetings, in particular, meeting browsers.

In the first phase of the research presented in this thesis, which was conducted in the context of the M4 project, no corpus was available to us that could be utilized for our study on addressing. Therefore, we developed a small multi-modal corpus of hand-annotated meeting dialogues designed primarily for studying addressing behavior in multi-party interaction. This was the first corpus to be made publicly available for this research purpose. Meetings included in the corpus were recorded in the IDIAP smart meeting room as a part of the M4 meeting data collection. These meetings are short, informal, discussion meetings. They 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. The spontaneous behavior of the participants in these meetings allows us to examine observable patterns of addressing behavior in small group discussions. The corpus developed in our study is a part of the publicly available M4 corpus which contains some additional annotations and meetings. In this thesis, we use the term M4 corpus only when we refer to our corpus. The AMI corpus, developed recently, consists of more natural scenario-based meetings involving a group focused on the design of a TV remote control. It was richly hand-annotated with a wide range of phenomena aiming to support multi-disciplinary research in the AMI project. Both corpora consist of four-participants meetings. By conducting experiments on both meeting corpora, we also aim to explore to what extent findings regarding the addressee identification in a less realistic meeting scenario can be generalized to more natural meetings.

The exploration of the relevance of various features is performed by means of static Bayesian Network classifiers. The features used to model addressing are obtained from human annotated data. As the features are in a sense ideal, the present study aims to investigate the upper bounds for the task of addressee prediction in face-to-face meetings. However, we also address the issue of the further automation of addressee identification process that could be obtained in a more realistic setting in which we have to rely on features provided by automatic predictors instead of human observers. Compared to static Bayesian Network classifiers, the Dynamic Bayesian Network classifier we consider in Chapter 5 is a step in the direction of this more realistic situation. This classifier uses the automatically identified value of the preceding addressee to predict the current addressee.

[1] the M4 project: http://www.m4project.org/

[2] the AMI project: http://www.amiproject.org/

1.4 Thesis outline

The remainder of this thesis is organized as follows:

Chapter 2 addresses the first research goal. It consists of two main parts. The first part is concerned with sequential organization of the conversational interaction whereas the second part deals with the organization of interaction across the participants. Furthermore, a clear distinction is made between conversation and meetings as two different contexts of talk. All phenomena under study are first introduced for ordinary conversation and then applied to meetings. The first part of the chapter describes two practices that locally organize conversational interaction - namely, turn-taking and adjacency pairs organizations - as well as overall organization of conversation. Since local organization practices were mostly studied in the two-party context, we discuss their limitations when applied to the multi-party case. This part concludes with an overview of various units of analysis that have been used by conversational and interaction analysts in studying different aspects of conversational interaction. Furthermore, it specifies the unit of analysis employed in our study. The second part of the chapter discusses the concept of participation in multi-party conversation. It first describes the notion of participant, different levels of participation and differences between official and unofficial participants in conversation. Then, several categorization systems are presented and compared. As outcomes of the analysis, the categorization of participation in our study is outlined and a definition of addressee is provided. The chapter further describes the recipient design principle focusing mainly on addressee design. It also discusses the relation among recipient design, turn-taking and adjacency pairs. Finally, the chapter describes a recently proposed framework for the analysis of conversational sequences that is based on the concept of participation in conversational interaction - the P-shift framework (Gibson, 2003).

Chapter 3 presents the M4 and AMI meeting corpora. The chapter starts with a brief description of the software that was employed for the creation of both meeting corpora - the NXT (NITE XML Toolkit). Each corpus is then described in detail including the description of the meeting data and annotation schemas. Regarding the AMI corpus, we focus only on those annotation schemas that are used for addressee annotation as well as for annotation of the features employed in our study. Apart from the corpora descriptions, this chapter provides a detailed reliability analysis of the annotation schemas. It concludes with a comparison between the M4 and AMI meeting corpora. The chapter is partially based on the work published in (Jovanovic et al., 2006b).

Chapter 4 provides a detailed overview of Bayesian Network models employed for the experiments on addressee identification presented in this thesis. It first introduces Bayesian Networks, well-known tools for knowledge representation and reasoning under uncertainty, and discusses various algorithms for learning and inference in Bayesian Networks. The second part of the chapter is concerned with Bayesian Networks in the function of classifiers. It provides an overview of the static Bayesian Network classifiers employed for automatic addressee modeling presented in the next chapter. Further, the chapter gives a brief description of Dynamic Bayesian Networks as they were also used for the same purpose. The chapter concludes with a discussion on methodologies for classifiers' evaluation.

Chapter 5 presents results on addressee identification in four participants face-to-face meetings using the static and dynamic Bayesian Network models introduced in the previous chapter. The first part of the chapter discusses the results of addressee classification on the M4 data performed by means of the static Bayesian Networks classifiers. The experiments can be seen as preliminary explorations of features and models for this type of task. The second part of the chapter is concerned with automatic addressee identification on the AMI data performed by means of the static and dynamic Bayesian Network classifiers. To compare performances of the static Bayesian Networks classifiers on the M4 and AMI data, the experiments on the AMI data were first performed using a set of features that were shown to be the most informative for addressee identification on the M4 data. The chapter further describes a number of experiments conducted on the AMI data using static Bayesian Network classifiers that were carried out to explore the relevance of additional features and models for addressee identification. The results of those experiments were employed in the design of a Dynamic Bayesian Network addressee classifier. The experiments with the Dynamic Bayesian Network classifier conclude the experimental part of this chapter. At the end, the chapter addresses the issue of further automation of the addressee classification using automatically extracted instead of manually annotated features. This chapter is partially based on the work published in (Jovanovic et al., 2006a).

Chapter 6 summarizes the thesis and suggests some general directions for future work.

Chapter 2

Addressing in the context of interaction research

To engage successfully in conversational interaction, participants are obliged not only to produce utterances, but also to coordinate their talk with the talk of others in a meaningful way. The coordination concerns both content and process of talk. In other words, participants have to successfully coordinate not only what they say, but also when they say what they say and to whom.

In this study, we are concerned with *the concept of participation* in conversational interaction taking place in a face to face setting. Moreover, we focus our investigation on meetings as a specific form of talk-in-interaction. As most analyses of phenomena implicated in conversational interaction have been conducted for ordinary conversation, we will first introduce the phenomena under study defined for conversation and then discuss their generalizations and adaptations for meetings.

The starting point of our study is Goffman's theory of the *participation framework* (Goffman, 1981a). Goffman claims that at any point in time, conversation can be characterized by its participation framework:

> When a word is spoken, all those who happen to be in perceptual range of the event will have some sort of participation status relative to it. The codification of these various positions and the normative specification of appropriate conduct within each provide an essential background for interaction analysis. (Goffman, 1981a, p. 3)

The participation framework represents an ensemble of 'participation statuses' of all participants in the gathering at a particular moment of speech. Goffman has also suggested that the traditional notion of speaker can be decomposed into a set of categories that he termed the *production format.* Developing this insight, Goffman was mostly interested in a decomposition of the global notions of hearer and speaker into a set of categories and in the way in which the boundary between official and unofficial participants in conversation is maintained. However, he did not provide us with sufficient characterization of the categories nor with clear distinctions between them to make their application feasible.

Another very important aspect of participation, that remained also unexplained by Goffman, is *recipient design.* It refers to a variety of methods speakers employ in designing their talk to designate which participants are to take which roles. The major contribution to the analysis of participation along this direction has been made by conversational analysts (Sacks et al., 1974; Sacks and Schegloff, 1979; Goodwin, 1981; Clark and Carlson, 1992; Lerner, 1995, 2003). According to Sacks et al. (1974) the recipient design represents "perhaps the most general principle which particularizes conversational interaction". Taking into account that talk-in-interaction, as has been discovered by conversational analysts, is organized by turn-taking practices and sequences of actions produced in interaction, a question that naturally arises at this point is "how are the recipient design principle and these two practices mutually influenced?". This chapter tries to provide an answer to this question by analyzing these three aspects both separately and in combination.

The first part of the chapter is concerned with the analysis of the phenomena that organize conversational interaction in a face-to-face setting. An intuitive definition of face-to-face conversation is provided in Section 2.1. Section 2.2 highlights some limitations of traditional approaches in studying conversational interaction based on the two-party model emphasizing the inapplicability of the outcomes of these studies on the multi-party model. An overview of the local organization practices operating in conversation - namely, turn-taking and adjacency pairs - as well as the overall organizations of conversation is given in Section 2.3. Section 2.4 introduces meetings as a special form of talk in interaction and provides comparisons between meetings and conversations regarding turn-taking design and overall organization. Section 2.5 summarizes various units of analysis that have been employed by conversational and interaction analysts in studying different aspects of conversational interaction. The second part of the chapter discusses the concept of participation from two perspectives: (1) categorization of participation, that is, the production format and the participation framework with a focus on the addressee role and (2) recipient design, with the a focus on the addressing process and addressing behavior. The concept of participation is first introduced for face-to-face conversation (Sections 2.6, 2.7 and 2.8) and then extended to meetings (Section 2.10). Relations between recipient design, turn-taking and adjacency pairs are described in Section 2.9. The last part of the chapter (Section 2.11) describes a new framework, the so-called P-shift framework, proposed in the literature for the analysis of conversational interaction which is concerned with moment-by-moment reshuffling of the participation framework.

2.1 Conversation - a preliminary definition

Conversation in a broad sense can be defined as a

> ..familiar predominant kind of talk in which two or more participants freely alternate in speaking, which generally occurs outside specific institutional settings like religious services, law courts, classrooms and the like (Levinson, 1983, p. 284)

This definition encompasses different kinds of conversations such as face-to-face conversations, phone conversations and computer mediated conversations. As the main focus of our study is a face-to-face setting, Goffman's definition of conversation can also be employed. According to Goffman (1981d, p. 14), the term conversation can be used "in a loose way as an equivalent of talk or spoken encounter" where spoken encounter can be seen as a specific kind of *focused interaction* (Goffman, 1963, p. 24):

> the kind of interaction that occurs when persons gather close together and openly cooperate to sustain a single focus of attention, typically by taking turns at talking.

On the other hand, *unfocused interaction* represents the kind of communication concerned with "management of sheer and mere copresence".

A more technical characterization of conversation will be provided in Section 2.4 after introducing three kinds of organization operating in conversation: turn-taking, adjacency pairs and overall structural organizations.

2.2 Dyadic versus Multi-party conversation

Traditional research on verbal interaction has been mainly focused on studying dyadic interaction. The term 'dyadic' is used to describe interaction between exactly two participants (the term 'two-party' is also employed). As such, in contrast to two-party interaction, multi-party interaction refers to interaction encompassing three or more participants. An alternative term used in the literature to describe multi-party case is 'polylogue' (Kerbrat-Orecchioni, 2004).

The main assumptions for privileging dyadic interaction was an implicit belief that the two-person arrangement "informs an underlying imagery we have about face-to-face interaction" (Goffman, 1981a, p. 129). It has also been assumed that without basic changes in analysis, any modification of conditions can be easily handled (e.g. the inclusion of additional participants in interaction). However, many researchers have shown that crucial phenomena regarding conversation discovered for two-party case, cannot be easily managed in multi-party case. Furthermore, many patterns of interaction found in interaction involving more than two participants are not applicable when only two participants are involved. For all phenomena discussed in this chapter, we will give some examples of cases that are not supported in a multi-party situation. Participation is an example of phenomena that is highly simplified in the dyadic model of conversation:

> The common dyadic model of speaker-hearer specifies sometimes too many, sometimes too few, sometimes too wrong participants. (Hymes, 1974, p. 54)

2.3 Conversational organization

A fundamental finding about talk-in-interaction in general, and thus about conversation in particular, is that talk is proceeded through a sequence of turns. *Turn* can be intuitively defined as “the talk of one party bounded by the talk of others” Goodwin (1981). This definition has a number of limitations regarding, for example, overlapping speech or backchanneling. A process through which turns are exchanged between speakers is referred to as *turn-taking.*

Conversational analysts, focusing their investigation on turn as a basic unit of analysis, have uncovered two fundamental forms of organization of conversation: organization of turn-taking and organization of sequences. Organization of turn-taking is considered with mechanisms that people use in interaction to achieve orderly and smooth transitions from one speaker to another. Although turns occur one after the other, in a serial order, relationship between them is not serial but sequential: each turn displays understanding of the previous turn. This principle of sequential organization is called *adjacency relationship* (Schegloff, 1988, p. 113). The adjacency relationship is the most general sequential organizational feature of conversation. It is a by-product of turn taking design: “it is a systematic consequence of turn-taking organization of conversation that it obliges its participants to display to each other, in a turn’s talk, their understanding of other’s talk”(Sacks et al., 1974, p. 728). A more complex sequential organization operating in conversations is *adjacency-pairs relationship.* Adjacency pairs are a kind of paired utterances such as greeting-greeting or question-answer. Compared to adjacency relationship, adjacency pairs have in addition “a powerful prospective operation”: the first paired utterance (termed as first pair part) limits a set of relevant second pair parts that come next (Schegloff and Sacks, 1973; Schegloff, 1988).

Turn-taking organization and adjacency pair organizations are two types of *local organization* operating in conversation - local in a sense that they are operating across just two turns (Levinson, 1983). Additionally, there are other types of organizations, called *overall organizations*, that organize conversation as whole i.e. the “totality of exchanges within some specific kind of conversation” (Levinson, 1983, p. 308-309). Each form of talk-in-interaction is characterized with its own form of overall organization.

2.3.1 Local organization of conversation

In this section, we discuss in more detail turn-taking and adjacency pair organizations of conversation. These two mechanisms are not the only organizational mechanisms operating in conversation. There are also quite different devices such as organization of repairs (Schegloff et al., 1977) or pre-sequences (Levinson, 1983). However, we describe only turn-taking and adjacency pairs, as we believe that they are directly related to the organization of talk across different participants.

The turn and turn-taking systems

Although the turn and turn-taking phenomena are intuitively clear and easily observable in conversation, providing a precise definition of a turn as well as specifying mechanisms that organize turn-taking is anything but simple (Goodwin, 1981; Levinson, 1983).

There is no general agreement among investigators as to what should be classified as a turn nor how turn-taking mechanisms operate in conversations.

Turn How turns are defined varies from study to study. Some definitions equate turns with speakers whereas others identify turns with isolated speech. For example, Goffman (1981d, p. 22-23) refers to a turn or a turn at talk as "an opportunity to hold the floor, not what is said while holding it", where "everything what an individual says during his exercise of turn at talk" is referred to as talk during a turn. On the other hand, Haris (1951, as cited in Goffman, 1981d, p. 22-23) defines a turn as "a stretch of talk, by one person before and after which there is a silence on the part of that person". The latter specifies the boundaries of a turn in terms of silence on the part of speaker. A slightly different approach includes other participants' behavior for identifying boundaries. For example, Fries (1952, as cited in Goodwin, 1981, p.17) defines a turn as "all the speech of one participant until the other begins to speak". Goodwin's (1981) definition of a turn introduced in Section 2.3 makes use of the same approach.

The cited definitions are primarily concerned with accurate determination of the boundaries of the turn. Problematic concepts for these kinds of definitions are *simultaneous speech* and *silence*. Simultaneous speech includes overlaps and interruptions. Regarding simultaneous speech, Fries (1952) definition runs into problems as a speaker's turn can be ended before he has finished pronouncing his utterance.

There is general agreement among researchers that silence should be classified as a "pause" if it occurs within a turn of a single speaker or as a "gap" if it occurs between turns of two speakers (Sacks et al., 1974; Goodwin, 1981). Conversation may also be discontinued which is manifested by a silence resulting after a current speaker has stopped and no other speaker starts or continues. These kind of silence is referred to as "lapse" (Sacks et al., 1974, p. 714). In terms of silence, the cited definition Haris (1951), for example, causes that pauses are treated as gaps. Furthermore, the same silence may lead to different classifications at different moments of speech and from the perspective of different participants (Goodwin, 1981).

< 1 > (*from (Goodwin, 1981, p.18)*)

```
1   John   Well I, I took this course
                 (0.5)
2   Ann    In h ow to quit?
               ]
3   John       which I really recommended
```

When Ann starts to talk, the silence between John's and her turns is considered as a gap as it is a silence between turns of two different speakers. However, John's talk a moment later is the continuation of the turn in progress that started before the silence. Since it is placed within the turn of a single speaker, it is considered as a pause.

Determination of turn boundaries is not only elusive for analysts but also for participants in conversation. As can be seen from the previous examples, turn-boundaries are changeable: at different points in talk, different boundaries can be assigned to the same turn. Thus, a definition of a turn as a static unit with accurate boundaries does not accurately describe the turn structure; it should rather be defined as a time-bounded process (Goodwin, 1981).

Approaching the problem from this perspective, many researchers defined the structural elements of turns i.e. the turn boundaries as implicated in the turn-taking process itself. Two solutions have been proposed as to how turn-taking actually works. One solution is that turn-taking is organized by a set of rules that operate on turn-by-turn basis (Sacks et al., 1974). The other solution is that turn-taking is regulated by a set of discrete signals (Duncan, 1972; Jaffe and Feldstein, 1970).

Rule-based turn-taking system Sacks et al. (1974) proposed a simple turn-taking system for conversation that consists of two components and a set of rules that operate on those components (henceforth, the SSJ system).

- **The turn-constructional component**- refers to the type of units out of which turns can be constructed. The key feature of such a unit is *projectability.* It denotes possibility that participants project, as the turn construction unit proceeds, what kind of a unit it is and what roughly will be needed to bring it to a possible completion. Such "unit-types" can be sentential, clausal, phrasal and lexical. Some turns are compounded of a single turn constructional unit while others of several.

  The property of projectability has been criticized in the literature (see O'Connell et al., 1990, p. 351). The main remark regarding Sacks et al. (1974)'s definition of turn-constructional unit is that the accomplishment of the projection is based solely on syntactic analysis. As has been discussed by Lerner (1996a), the term turn-constructional unit merely registers the fact that there are some units of which turns can be constructed. However, it does not imply that identification of a unit type is purely syntactical. Participants in conversation can rely on turn position (e.g. sequential environment) as well as turn composition (e.g intonation, structure, content) to identify which type of unit it is and thus, what will be needed to reach a possible completion of the turn (Lerner, 1996a, p. 306).

  The projectability has two main consequences. First, the end of a turn-constructional unit specifies a point where a speaker change can occur. It is termed a *transition-relevant place* (TRP). Second, the projectability specifies the limits of the speaker's current right to talk: the speaker is initially entitled, in having a turn, to one such unit; at TRP speaker transition becomes relevant (Goodwin, 1981).

- **The turn-allocation component**- The turn-allocation components consist of two groups of allocation techniques:

  1. **'current-selects-next'** techniques are techniques in which the next turn is distributed by the current speaker selecting a next speaker. An example of such techniques is the use of the first-pair part of an adjacency pair (e.g. question)

that employs certain addressing practices such as the affiliation of an address term (e.g. a personal name or a categorical term of address (e.g. 'teacher')). For example, the question *Ann, are you coming tomorrow?* not only constrains that in a next turn an answer should be provided, but also selects Ann as the next speaker who should provide that answer.

2. **'self-selection'** techniques are techniques in which the next turn is distributed by self-selection of participants other than the current speaker. The basic technique of this kind is 'starting first'. It denotes that after a brief gap, only one speaker starts. "The *single* starter should be thought of as 'first starter' succeeding in being single because of the 'first starter' provision [incorporated in Rule 1b, forthcoming], and being 'dispensable' in that, if he had not started fast, someone else would have" (Sacks et al., 1974, p. 719). Being pressured to start as early as possible in competing for next turn, a self-selector would have to start in many cases while the current turn is still in progress, close to its TRP.

- **Rules** - The following set of rules governs operation of turn-taking in conversation.

  1. Rule 1 operates at an initial TRP of any turn.
     (a) if a current-selects-next technique is used in the current turn, then the selected speaker has the right and is obliged to take the next turn. The transition occurs at the initial TRP.
     (b) if such an allocation technique has not been used, then self-selection for next speakership is allowed, but not required, at this place; first starter gets rights to next turn and transition occurs at this place.
     (c) If a current-selects-next technique has not been used nor another self select, the current speaker may, but need not, continue.
  2. Rule 2 operates at all other TRPs

     If current speaker continues into a new turn-constructional unit applying Rule 1c, then Rules 1a-1c reapply at the next TRP and all subsequent TRPs until a transition to a new speaker occurs.

Properties of the SSJ turn-taking system The proposed system has the following properties: (1) it is *locally managed* as it deals with a single transition at a time, organizing thereby only two turns - current and next; (2) it is *party-administrated* as the control over its operations and products is subjected to participants in conversation; and (3) it is *interactionally managed* as constructions of turns and determination of their boundaries involve tasks distributed among participants (Sacks et al., 1974).

The turn-taking rule set provides an *ordering* of applications of two groups of turn-allocation techniques which supports the '*one speaker at a time*' characteristic of conversation. For example, if both group of techniques could be applied on any occasion they would permit more than one speaker to be selected (Sacks et al., 1974). Claiming this, however, Sacks et al. (1974) did not consider some instances of 'current-selects next' techniques when more than one speaker is selected, and thus both groups of techniques, regardless of their pre-specified ordering, may be applied on the same occasion effecting simultaneous talk. We will elaborate on this later when we discuss the limitations of the turn-taking system.

The turn-taking rules also support localization and minimization of *overlap*. As transitions between speakers occur at TRPs, overlaps in a large number of cases occur at TRPs and their immediate environment. This is one reason why overlaps are brief. There are several systematic bases for overlaps. Overlaps may occur either (1) when several self-starters compete for a next turn, when each of them projects his start to be earliest possible at some TRP or (2) when TRPs are incorrectly projected due to systematic reasons. For example, the current speaker may add an addressed term after the first possible completion without intending continuation as he has already applied a 'current-selects-next' technique (Sacks et al., 1974).

The turn-taking rules also describe some properties of the structure of the turn itself. A turn is not defined as a static unit but through a time-bounded process with discrete but alterable boundaries (the application of Rule 2). These rules also enable different classifications of silences. For example, the silence after the application of Rule 1a can be classified as a gap. Furthermore, the rules enables transformation of one type of silence into another one. For example, if a developing silence occurs at a TRP, and is thus a potential gap, it may be ended by the talk of the speaker of the previous turn-constructional unit, becoming thus a pause (Sacks et al., 1974, p. 715, f. 26).

Sacks et al. (1974, p. 699) state that the proposed turn-taking organization for conversation is a 'context-free' organization that is capable of 'context-sensitivity'. The term 'context' is employed to refer to various places, times and identities of parties in interaction. The system is context-free in the sense that it is insensitive to these parameters of the context. Although the system is context-free, it defines how and where context-sensitivity can be displayed (i.e. it shapes particularities of the context). For example, some current-selects-next techniques employ addressing practices that are context-sensitive (see Section 2.7.2).

Some limitations of the SSJ turn-taking system for conversation Analyzing conversations among Americans, Sacks et al. (1974) observed, among other properties, that (1) most of the time only one speaker talks, (2) occurrences of more than one speaker are common but brief and (3) smooth transitions, with no gap and no overlap or with slight gap and slight overlap, constitutes a vast majority of transitions and (4) the number of parties in conversation can vary. The SSJ model is designed to support all these characteristics for conversation. It is supposed to be a context-free system in the sense that it is insensitive to places, times and individuals involved in interaction. However, many researchers run into problems trying to apply the formal SSJ model on the actual data. The reasons for this are threefold:

1. conversational conditions assumed in the system are oversimplified

2. the system's components and rules are simplified (see O'Connell et al., 1990, 360)

3. the varieties of turn-taking situations that occur in multi-party cases have not been investigated at all

There are several reasons for occurrences of overlap and interruptions in conversation. First, occurrences of overlap and interruptions are influenced by culturally determined conversational style. In American culture, speakers talk in turns (O'Connell et al., 1990). As mentioned above, the SSJ model is based on an underlying rule of an American English conversations that mostly one speaker talks at time. In contrast, some other cultures have preference for simultaneous speech. Italians, for example, tend to interrupt one another as "everybody gets all excited and tries to make his views prevail by preventing the other from speaking" ((Eco, 1986) as cited in (O'Connell et al., 1990)). Second, overlap and interruptions are influenced by participants emotional state and relationship among them. It is to be expected that an agonistic relationship leads rather to an overlap than to smooth turn-taking (O'Connell et al., 1990). All of these elements are part of conversational settings and must be taken into account while analyzing conversation. Some analyses of simultaneous talk in conversation show that the frequency of overlaps and interruptions significantly increases when more than two participants are involved in interaction (Shriberg et al., 2001; Kerbrat-Orecchioni, 2004).

Not all instances of overlap and interruptions are always degenerate, brief, or something that need to be repaired. On the contrary, they can be functional and acceptable for the current speaker. Lerner (1996a) identified some cases of turn constructions where a turn of one speaker can be completed by another speaker - *the anticipatory completion.* Investigating complex turn constructional units that are constructed out of a multi-clausal sentence, Lerner (1991) defined *compound TCU* that projects completion in a different way. The compound turn constructional unit has a multi-component structure that consists of at least a preliminary and a final component. Some common types of compound TCU include "if X + then Y" or "when X + then Y". According to the SSJ system, a compound TCU can reach its completion only after the final component is produced. Therefore, participants must examine the emerging talk to determine whether the final component has begun. Only after the final component has begun, will participants examine the proceeding talk to search for the upcoming TRP. As in the course of the preliminary component participants can project where the final component could start and what form the final component will take, the turn can be completed by another speaker. Turn sequences produced in such a way, are called *collaborative turn sequences.* According to Lerner (1996a), the collaborative turn sequences imply a relaxation and not violation of the SSJ turn-taking system.

Multi-party interaction features violations of speaker-selects-next rules. Applications of Rule 1a assume that the selected speaker is required and obliged to take the next turn. However, it multi-party interaction it can happen that the participant who takes the turn is not the one who is selected by the current speaker. One example of such interaction patterns is *piggybacking* (Goodwin and Goodwin, 1990). Piggybacking is a form of interaction when someone "readdresses the prior target as a way of affiliating with prior speaker" as it is presented in the following example:

< 2 >

1 A: Are you coming with us?
⟹ 2 B: to the cinema
3 C: Yes sure.

In multi-party conversation, a speaker can also use current-selects-next techniques to address more than one participants. For example, *Ann and John, where did you go yesterday?*. If more than one participant is addressed, the SSJ system does not provide for a selection of a person among the addressed participants to speak next. The reason for this is that the SSJ turn-taking model organizes distribution of talk not between participants but between parties (Schegloff, 1995). Parties can be composed of more than one person. Multi-participant parties may coincide with units of social organization such as couples or classmates. Furthermore, momentary associations of participants into a party is also possible during the course of interaction (Lerner, 1993).

Signal-based turn-taking systems Another view to turn-taking organization is that turn-taking is regulated, by a set of discrete signals (Jaffe and Feldstein, 1970; Duncan, 1972). According to a signal-based turn-taking system, the current speaker will provide a signal when he wants to relinquish the floor - 'end-of-the message' signal (Jaffe and Feldstein, 1970), 'turn-yielding-signal' (Duncan, 1972). However, the other participant is not required to speak after one or more of these signals are displayed.

Based on empirical observations of dyadic conversations, Duncan (1972) described a set of six turn-yielding signals: rising or failing pitch at the end of the clause, drawl, termination of any hand gesticulation during a speaking turn, use of stereotyped expressions such as *you know* or *or something*, a drop in pitch or loudness in combination with stereotyped expressions and the completion of a grammatical clause. Regardless of the number of turn-yielding signals displayed, the speaker can neutralize them by displaying an 'attempt-suppressing-signal', also called 'gesticulation signal' (Duncan and Niederehe, 1974). This signal consists of the speaker maintaining gesticulation of his hands during a turn-yielding signal. According to empirical analysis presented in (Duncan, 1972), participants almost never attempt to take a turn when attempt-suppressing-signal has been displayed. Furthermore, it is more likely that a participant will take a turn if more turn-yielding signals are displayed at a specific moment.

The presented signal-based system has several disadvantages in comparison to the SSJ turn-taking system. As discussed in (Goodwin, 1981) and (Levinson, 1983), it is not clear how signal-based mechanisms regulate some observable patterns of turn-taking such as lapses and gaps. Duncan (1972, p.286) states that avoidance of the turn by both participants was not present in the data he used in analysis. As the system is limited to termination points of turn-constructional units, it does not make a clear distinction between current-selects-next and self-selection techniques. Furthermore, the SSJ system provides some systematic possibilities of overlap at TRP, whereas for Duncan (1972, p. 286), overlap means that "the turn-taking mechanisms may be said to have broken down" . As the signal-based turn-taking system does not support many phenomena that characterize conversation, it would not easily accomplish the turn-taking mechanisms in multi-party interaction.

Turn vs. Floor In most of the cited work on turn-taking, the terms turn and floor are used interchangeably. Furthermore, the cited definitions of the turn are based mostly on speaker exchange. This technical approach in defining turn does not take into account either a participant's intuition as to what constitutes turn or the speaker's intention (Edelsky, 1981). Approaching the problem from this perspective, Edelsky (1981) makes a clear distinction between turns and floors.

Turn is according to Edelsky defined as "an on-record "speaking" (which may include nonverbal activities) behind which lies an intention to convey meaning and to convey a message that is both referential and functional."

Many researchers agree that not all talk is considered as a turn. Common off-records are *side comments* that are not meant for public hearing. Produced usually with a lower voice tone, they are intended to be observed by a single participant or a subgroup of participants. The definition of turn also excludes those utterances in which a speaker intends to provide only a feedback but not a referential message (e.g. *yeah*, *hm*, *okay*). In many cases, they are actually encouragements to the speaker to continue talking. Aborted utterances that do not convey a part of a referential message are not considered as turns either (e.g *but I*). The definition of turn requires also that the speaker's intention is taken into account in determining turn-boundaries as it should convey a functional message. Edelsky's definition of turn is adopted in this thesis for developing models for automatic detection of who is being addressed by the current speaker.

Edelsky defines a floor as the "the acknowledged what's-going-on within a psychological time and space". What's-going-on can be, for example, the development of a topic or an interaction between a subgroup of participants. It is acknowledged if participants can describe what is going on as *he is talking about the remote control* or *we are agreeing with her.* A floor can, therefore, consist of several turns. On the other hand, it is possible to take a turn without having the floor. A person may also hold the floor while not talking. For example, while explaining a complicated formula, a speaker can write down the formula without saying anything for some period of time. The floor can be controlled by one person (*single developed floor*) or by several persons simultaneously or in succession (*group developed floor*).

Adjacency pairs

In addition to turn-taking organization, conversation is locally organized by adjacency pairs. Adjacency pairs, as defined in (Schegloff and Sacks, 1973), are sequences of two utterances that have the following properties:

- adjacent positioning of component utterances
- each utterance is produced by a different speaker

- utterances are relatively ordered as 'first pair part' and 'second pair part'; first pair parts precede second pair parts

- a particular first pair part requires and constrains a range of possible second pair parts

Prototypical instances of adjacency pairs include 'question-answer','greeting-greeting' and 'offer-accept/refusal'.

The properties that characterized adjacency pairs are inappropriate in two aspects (Levinson, 1983). First, the requirement for the strict adjacency is too strong. Second, the range of acceptable second-parts for a given first part do not constitute a small set.

Very often in conversation one adjacency pair can be embedded within another adjacency pair forming the following sequence $\mathrm{A}_1(\mathrm{A}_2\text{-}\mathrm{B}_2)\mathrm{B}_1$, where A*i* and B*i* represent the first and second part of the i-th adjacency pair, respectively. The embedded adjacency pair is called an *insertion sequence.* Insertion sequences can be nested. In the following example one question-answer pair is embedded within another:

< 3 >
1 C: Where did you buy your jacket? (A1)
2 D: The black one? (A2)
3 C: Yes (B2)
4 D: In Italy (B1)

In multi-party conversations, even more, 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, an offer can be followed by either acceptances or refusals from multiple speakers. Piggybacking is also an example of violation of the strict adjacency condition in interactions involving more than two participants (Goodwin and Goodwin, 1990): a speaker insert his talk within an adjacency pair environment produced by another speakers by affiliating his talk to the first pair part and readdressing the addressee of that part.

The second problem regarding adjacency pair conditions is the range of acceptable second parts for a particular first part (Levinson, 1983). For example, there are many responses to a question other than answers that can be viewed as acceptable second parts including refusals to provide an answer, postponements, statements of ignorance, suggestions for asking someone else or denials of the relevance of the question. However, not all second parts relevant to a given first part are of equal standing. There can be some second parts that are *preferred* and some that are *dispreferred.* For example, if requests or offers are realized by the first part, acceptance is the preferred and the refusal is the dispreferred second part. This ranking of second parts as preferred or disprefered is encompassed in the concept of *preference organization*(Levinson, 1983).

It is to be noted that not all utterances in conversations are marked as a first or second pair part of an adjacency pair. Furthermore, there are also sequences produced in conversation that are not characterized as adjacency pairs (e.g. story-telling sequences (Sacks, 1992)).

2.3.2 Overall organization of conversation

Overall organizations organize the totality of exchanges within some specific kind of conversation in a structural way. There are three parts of the overall organization of the unit 'a single conversation': opening, topical structure and closing (Schegloff and Sacks, 1973; Schegloff, 2002). These aspects of overall organization have been extensively studied in the context of phone conversations (Schegloff and Sacks, 1973; Schegloff, 2002, 1986; Levinson, 1983).

Openings of phone conversations are constructed from the following four types of opening sequences (Schegloff, 1986): (1) the summons-answer sequence - (*phone rings-Hello*), (2) the identification sequence possibly followed by the recognition sequence - receiving and calling party identify each other and display each other's recognition of the other, (3) the greeting-greeting sequence (*Hi-Hi*) and (4) the *how are you* sequence - (*How are you?- Fine. And you?*). After the ending of the opening section, the calling party announces the reason for the call which is referred to as '**first topic**' (Schegloff and Sacks, 1973, p: 300-301). The first topic is a topic that can be characterized by participants as the reason for conversation. As such, it is a "preservable and reportable feature of the conversation". The announcement of the first topic is followed by topic closing and shifting to next topic. Finally, a topical sequence is followed by a **closing section** that may consist of elements such as closing down of some topic, exchange of turns such as *Okay* or *All right*, identifying the type of call (e.g. *I just call to check how you feel*) followed by a further exchange of messages and the terminal exchanges of elements such as *Bye* or *See you* (Levinson, 1983).

It is to be noted that the term topical organization can refer both to the organization of the unit 'a topic' and to the organization of a set of such units within the unit 'a single conversation'. Though closing down the topic can be relevant for closing conversation as its initiator, the topical organization as an aspect of the overall organization is mostly considered with the latter.

Opening and closing sections as well as topical organization as described in this section are applicable with some modifications to other similar kinds of conversation such as conversation during a chance meeting on the street or internet chatting.

2.4 Conversations and meetings - two forms of talk-in-interaction

To provide more technical definition of conversation, Levinson (1983) makes a distinction between *conversational activity* and the unit *'a conversation'*. Conversational activity is characterized in terms of local organizations that operate on turn-by-turn bases - especially in terms of turn-taking systems. There are, however, many forms of talk-in-interaction that cannot be characterized as conversational activities. Examples include lessons and sermons, where a lecturer or a preacher speaks with little or no opportunities for interruption by members in the audience. On the other hand, there are some forms of talk-in-interaction that have properties of conversational activities - such as courtroom or classroom interrogation - but that are clearly not conversations. According to Levinson (1983), conversation as a unit is characterized not only by the use of conversational activities such as turn-taking of the kind presented in Section 2.3.1, but also by the overall organization of the sort presented in Section 2.3.2.

Meetings are usually devoted to business and scientific or political matters. They can be highly formal such as chaired city council meetings on the one hand, or informal such as some brainstorming meetings of working groups on the other hand. Participants in a meeting may have been assigned certain roles regarding the meeting as a whole (e.g chairmen) or regarding particular activity type taking place in a meeting (e.g. presenter, demonstrator). Meetings are usually planned in advance. In many cases, especially in formal settings, there is an agenda setting up the phases and activities that the meeting will contain and topics that will be discussed. Some highly formal meetings can also strongly constrain participants' turn-taking behavior such that participants have to rise a hand to ask for the word and that the chairman is the one who allocates the next speaker. In contrast, informal meetings pose less restrictions on participants' behavior of this kind.

Conversations and meetings as two forms of talk-in-interaction may differ across several dimensions. In this section, we focus only on turn-taking design and overall organizations.

The turn-taking system for conversation, as presented in (Sacks et al., 1974), supports the *local allocation* of turns. Who talks and what is talked about is decided on turn-by-turn basis, by the participants in the conversation. On the other hand, meetings may pose some restrictions to the turn-taking design. In meetings with chair persons, turns are *partially pre-allocated*. The chair person is authorized to control the turn-taking: he has the right to talk first, to talk after each speaker and to use such turns to select a next speaker (Sacks et al., 1974, p. 729). It is to be noted that in chaired meetings not all turns are pre-allocated. The participants can be engaged in a discussion that resembles ordinary conversation in a sense that turns at speaking are locally allocated. Investigating the organization of turn-taking in a chaired political meeting, Larrue and Trognon (1992) observed that the turn-taking system for that meeting contained the SSJ turn-taking system for conversation. The turn-taking design in meetings can be restricted in other ways as well. For example, a presenter during the presentation can take over a leading role for that part of the meeting coordinating the turn-taking process in a similar way as a chairman. As noted in (Clark, 1996), turns in informal business meetings may be allocated by explicit agreement among participants. Participants may agree upon who is to talk on what and for how long. The distinction in local allocation and partial pre-allocation of turns effects the difference in the size of turns allocated in such way: "turn-size increases with increasing the degree of pre-allocation" (Sacks et al., 1974, p. 730).

Meetings also differ from conversations in the way they are structurally organized. Meetings can be formally structured according to the agenda. During the meetings participants can perform various types of group activities that are not common for conversation such as giving a presentation or a demonstration or voting for or against a proposal. Opening and closing of meetings are activities more complex compared to the corresponding activities of conversation. Opening a meeting may consists of taking attendance by a chairman ("*Hello everybody*"), presenting or setting up an agenda, discussing the time schedule, presenting goals of the meeting. Similarly, closing may consists of planning the next meeting by setting the date and distributing of the tasks among participants for the next meeting, and of terminal exchanges at the end of the meeting. Structural organizations of meetings may have an influence on participation frameworks as we shall discuss in Section 2.10.

2.5 Units of analysis of conversational interaction

When participants are engaged in spoken interaction, what they produce with their talk are *utterances* rather than *sentences*. The term 'sentence' is often used to refer to written form. Bloomfield (1946) defines sentences as "an independent linguistic form, not included by virtue of any grammatical construction in any larger linguistic form" (cited in Goodwin (1981)). Sentences are abstracted away from the context that refers to particular speakers, listeners, times and places. On the other hand, utterances are the result of producing words on a particular occasion, by a particular speaker for a particular purpose (Clark, 1996).

Similar to the disagreement among investigators as to what should be defined as a turn, there is no general agreement regarding the definition of an utterance. Some researchers equate turns and utterances as a result of defining turns based on isolated speech (Haris, 1951). Goodwin (1981, p.7) defines an utterance as an "actual stream of speech actually produced by a speaker in conversation". It includes "the entire vocal production of the speaker - that is, not only those sounds that could be placed in correspondence with elements of sentences, but also phenomena such as midword plosives, inbreaths, laughter, crying, "uh's" and pauses". Goodwin (1981, p.7-8), however, does not separate an utterance into subunits that have sentence-like properties allowing the possibility that an utterance consists of several sentences as well as the possibility that a sentence is compounded of several utterances. Goodwin's definition of utterance is used in this study. It is to be noted, that not all utterances are considered as turns. Backchannels, side comments and aborted speech are examples of utterances that are not turns (see Section 2.3.1).

Producing their talk, speakers intend not only to *convey a message* but also *to perform certain actions toward the particular others* such as asking a question or suggesting an action to be taken. This aspect of speaker intention should be taken into account for interactional analysis of conversation. What we are concerned with is not sequences of turns per se but sequences of actions produced in talk. As Goffman (1981d) has noted, utterances cannot be used as a basic unit of interactional analysis because of the following reasons:

1. with an utterance a speaker may perform several different actions (e.g. *yes I will come* ||*but what about you?*).

2. talk during two different turns can function as one interactional unit. For example, a question may be shared by two different persons as in the following example:

< 4 >

1	A to C:	Do you have time?
2	B to C:	to visit us tomorrow
3	C to A and B:	Yes sure

3. within one utterance two different parties' contributions can be combined. A speaker may step in to help the pervious speaker to find a word that he was looking for to complete a question and then immediately provide a response. In this case, within an utterance contributions of the previous and the current speakers are combined.

To overcome the listed problems, Goffman (1981d) suggested the use of a notion of 'move', although without providing a precise definition of that notion. In a broad sense, a move refers to

> "any full stretch of talk or of its substitutes which has a distinctive unitary bearing on some set or other of circumstances in which participants find themselves ... such as a communication system, ritual constraints, economic negotiating, character contests,"teaching cycles" (...), or whatever." (Goffman, 1981d, p. 24)

From this, it follows that an utterance that is a move in one game may also be a move in another, or be a part of a move of another game or contain two or more moves of another game (Goffman, 1981d). In the following example, B's response can be seen as a move in three different games: (1) the requested information is provided, (2) the question was correctly heard and (3) it was not impolite to ask such question (Goffman, 1981d):

< 5 >

A: Is this the book you advised me to read?
B: Yes

Goffman's notion of interactional move encompasses different types of "doings" that participants perform in conversational interaction within certain occasions. One dimension along which the interactional moves can be classified is the *speech act* dimension that refers to a number of acts that are performed by speaking (Austin, 1962; Searle, 1969).

2.5.1 Speech acts

According to Austin (1962), with each utterance a speaker is performing three types of acts:

- **locutionary act**: uttering a sentence with a certain meaning.
- **illocutionary act**: uttering a sentence that has a certain *conventional* force e.g. asking, informing or offering.
- **perlocutionary act**: uttering a sentence that produces certain effects upon feelings, thoughts, attitudes or actions of the addressee such as misleading, insulting, convincing or forcing.

However, the term speech act is used mostly to describe illocutionary act. Searle (1969) suggested the following five classes of speech acts (as described in Levinson (1983)):

- **representatives** - committing the speaker to the truth of the expressed proposition (e.g. asserting, concluding or boasting).
- **directives** - attempts by the speaker to get the addressee to do something, namely, to carry out a course of actions expressed by the propositional content (e.g. requesting, questioning, advising or ordering).
- **commissives** - committing the speaker to some future course of actions (e.g. offering, promising, planning or threatening).
- **expressives** - expressing a psychological state of the speaker (e.g. thanking, welcoming or apologizing).
- **declarations** - effecting an immediate change in the current state of affairs in the world (e.g. declaring a war, firing from employment).

Searle's taxonomy of speech acts is based on the *illocutionary force* which determines a goal and a reason for communication. A speaker actually intends to produce a certain illocutionary effect by means of getting the hearer to recognize his intention to produce that effect. However, the relation between the surface form of an utterance and its underlying purpose is not always straightforward. With an utterance, a speaker may perform one speech act indirectly by performing another. This indirectly performed act is called *indirect speech act* (Searle, 1975). Indirect speech acts are performed by the means of direct ones aimed at the same speaker. For example, the utterance *Can you open the window?* is expressed in a form of question. Usually, the main purpose of a question is to get an answer. This kind of illocutionary force that is incorporated into sentence form is called the *literal force* (Levinson, 1983). However, the main purpose of this utterance is requesting the addressee to open the window, which represents an inferred *indirect force.* Indirect force clearly dominates over literal force.

The problem of indirect speech acts is not the only difficulty in form-to-force mapping between utterance or utterance units and speech acts. An additional problem is that the same utterance can have different forces within different conversational contexts, as it is illustrated in the following examples:

A: What time is it?
B: five o'clock
A: Thanks.

A: We shall meet tomorrow
at five o'clock at the same place
B: five o'clock?
A: yes, five o'clock

A: We can meet later
B: five o'clock
A: Great. See you then
at five o'clock

B's utterance *five o'clock* has been used for a different purpose in all three examples. In the first example, B's utterance is an answer to A's question. In the second example, B's utterance reveals the speaker's intention to verify whether he understood or heard correctly what the previous speaker had just said. The main purpose of B's utterance in the last example is to suggest to the other participant that they meet at five o'clock.

The traditional speech act theory fails to provide for more than two illocutionary forces - literal and indirect, as an utterance is observed in isolation of the conversational context. One solution to a pragmatic theory of speech act has been proposed as the *context-changed theory of speech acts* (see Levinson, 1983, p. 276). The basic intuition can be explained as follows: "when the sentence is uttered more has taken place than merely the expression of its meaning; in addition, the set of background assumptions has been altered. The contribution that an utterance makes to this change in the context is its speech act force or potential". It is to be noted that Goffman's definition of moves also highlights the importance of the context.

A similar rationale has been explored by computational linguists in more recent work on dialogue modelling where a notion of *dialogue act* has been introduced. The term *dialogue* is used as a synonym for conversation. As the dialogue act is used as the basic unit of analysis in this study, it will be introduced in more detail in the following section.

2.5.2 Dialogue acts and interactional moves

According to Bunt (2000, p.5), dialogue acts can be defined as "functional units used by the speaker to change the context". A dialogue act has a 'semantic content' and a 'communicative function'. Semantic content represents certain information that is of particular importance for the new context. Communicative function defines the importance of this information by "specifying in what way the context should be updated to take this information into account" Bunt (2000, p. 5-6). In the literature, the terms *conversational moves* (Carletta et al., 1997) or *conversation acts* (Traum and Hinkelman, 1992) are used as an alternative for dialogue acts. The definition of dialogue act by itself does not assume that dialogue acts are of verbal character only. Shaking hands or nodding are also functionals unit used by the speaker to change the context.

Speech acts are just one aspect of communicative function. The notion of dialogue act covers the different kinds of functions that an utterance can fulfill in conversation. Popescu-Belis (2005) discusses the following seven dimensions of communicative function: speech acts, turn management (e.g. floor grabber or floor holder), adjacency pairs (e.g. question-answer), topical organization (e.g. opening, closing or topic-continuer), politeness (e.g. encouragement, self-depreciation or neutral) and rhetorical relations (e.g. elaboration, evaluation or purpose). These dimensions are not mutually exclusive. For example, an utterance may be a question as a speech act and as the first part of an adjacency pair. Furthermore, an utterance can be *multi-functional* in a sense that it has various functions across different dimensions.

In last years, there has been an ongoing effort to develop *dialogue act taxonomies* that enumerates the possible functions that an utterance can fulfill in dialogue - dialogue act types. An overview of several widely used dialogue act taxonomies is presented in (Popescu-Belis, 2005). In Chapter 3, we will describe in more detail two dialogue act taxonomies employed in the designing of the M4 and AMI meeting corpora.

It is to be noted that the notion of interactional move introduced by Goffman (1981d) is more similar to the notion of dialogue act than to the notion of speech act. Although the notion of dialogue act is used as the basic unit of analysis in this study, the terms move or interactional move will be used in the remainder of this chapter for the explanation of the phenomena that are to be introduced. The term dialogue act will be employed starting from Chapter 3.

Introducing the move as the basic unit of interactional analysis requires a revision of the notion of adjacency pairs in a way that the first and second parts of adjacency pairs refer to moves instead of to utterances. More generally, we can say that the conversation is organized as a sequence of moves rather than as a sequence of turns.

2.5.3 Utterance-events and interactional moves

In addition to interactional moves, which segment an utterance into utterance units that have certain communicative functions, Levinson (1987) introduced the notion of 'utterance-events' that segment an utterance into utterance units based on the changes in the participation framework. According to Levinson (1987, p. 168), an 'utterance-event' is defined as "that stretch of talk over which there is a constant set of participants' roles mapped into the same set of individuals - i.e. that unit within which the function from the set of participant roles to the set of individuals is held constant". Therefore, utterances can be segmented into utterance units based on two criteria: communicative function and participation framework. In some cases, the move boundaries may coincidence with the utterance-event boundaries whereas in other cases, moves and utterance-events may overlap. The following examples of conversations among three participants illustrate some of the possible relations between utterance-events and moves:

1. John: I agree with you || but what shall we do about that?
   two moves of different types (agreement || question); two utterance-events (addressed to the previous speaker || addressed to the group)

2. John: Ann, give me the book || and you, Mike, give me the pen.
   two moves of the same type (request || request) that can be marked as one move as well; two utterance-events (addressed to Ann || addressed to Mike)

3. John: Mike, please open the window || it is too warm.
   two moves (request || assertive); one utterance-event (addressed to Mark)

These examples indicate that the notion of move and the notion of utterance-event in multi-party conversations should be combined. In other words, what characterized a functional unit i.e. a move is not only its communicative function but also its participation framework. Two requests in the second example are clearly two different types of moves as they change the context in different ways by influencing actions of the different participants.

2.6 Participation roles in conversation - production format and participation framework

Goffman (1981b, p.3) uses the term '*participant*' to refer to any individual who is "in perceptual range of the event". As the 'perceptual range' involves visual and auditory co-presence of interacting participants, the term does not encompass other types of communication for which either of these two channels is excluded, such as phone conversations or written discourse. On the other hand, the term is too broad as it includes not only official participants in interaction but every person who is present in the situation. When persons are mutually present, they all or some of them can be communicated merely by virtue of their copresence and therefore can be considered as participants in interaction. Communication channels include, among others, gestures, emotional expressions, movements, positions, clothing and bearing. Goffman's notion of participant applies to conversation, or some other form of talk-in-interaction, as a whole. Defining participation at the level of conversation, Goodwin (1981), opposite to Goffman, defines as participants only those individuals that are engaged in a conversation. Someone not part of relevant conversation is referred to as a '*nonparticipant*'. Highlighting that the distinction between participants and nonparticipants in some cases may be unclear and ambiguous, Goodwin suggested that the notion of participant should be used in a broad enough sense to include those individuals who are momentarily disengaged. Levinson (1987), however, defines the term 'participant' at the level of conversation as well as at the level of utterance-event. At the level of conversation, only ratified participants, in Goffman's sense of ratification, are considered to be participants. The notion of ratification will be introduced in Section 2.6.2. It is possible to be a participant even being non-attentive or momentarily not being in audible range of the speaker. However, at the level of utterance-event, only those who are ratified and attentive are considered as participant. Similar to Levinson, Clark and Carlson (1992) define participants on the level of illocutionary act performed by a speaker, as those hearers who are intended by the speaker to take part in the illocutionary act he is performing towards the addressees.

In contrast to approaches mentioned above where participation is seen as a product of collaborative actions between speaker and hearers, Clark and Carlson (1992) describe participation in terms of speaker's activities: the speaker "defines his own role as speaker; he defines who is to "take part in" his illocutionary act". Taking account, first and foremost, that turn-taking system is interactionally managed, Clark and Carlson (1992)'s approach can be considered as oversimplified. Furthermore, many researchers have shown that the audience should be treated as active participants in conversation so that most aspects of interaction, including participation, are truly interactive and not merely outcomes of speaker activities projected in a field of participants (Goodwin, 1981; Duranti, 1986; Haviland, 1986; Goodwin, 1986).

2.6.1 Levels of participation

In defining participants in conversation, three different levels of analysis can be considered (Goodwin, 1981):

- **the activity of conversation** - The two most important positions for participants provided by the activity of conversation are *speaker* and *hearer*. These two positions are related to the activities of speaking and being silent.

  Goffman (1981a) uses the term ‘speaker’ to refer to an individual active in the role of utterance production - ‘the talking machine’. Some researchers define the term ‘speaker’ in terms of turn. For example, Goodwin (1981) refers to a speaker as a “party whose turn is in progress at a particular moment of time”. As pauses may occur within a turn, a party can be a speaker although he is not saying anything at the moment. However, simultaneous talk may pose problems for this definition. The term ‘hearer’, at this level of analysis, is used to refer to a complementary position to the position of speaker - a ‘non-speaking’ participant.

- **individual actions** - Distinct from the position of speaker and hearer provided by the activity of conversation are the actions of individuals that display their incumbency in these positions. Participants, for example, may not be listening while being in the position of hearer or even more so while being addressed. Goodwin (1981) gives an example of the situation where A is addressing an utterance to B who is, however, paying attention to another speaker C. The proper description of A’s action includes B as the addressee regardless whether B displays hearership to A. On the other hand, the description of actions of B relevant to the position of hearer includes C as a speaker he is being attentive to.

- **collaborative actions** - As mentioned above, many aspects of conversational interactions are truly interactive in a sense that they include actions of more than one individual. The process of ratification (Section 2.6.2) and the process of designing a talk for different types of participants (Section 2.7) are defined at this level of analysis. For example, speakers’s design of a talk for a particular participant in conversation is influenced by the orientation and attentiveness of that participant towards the speaker.

2.6.2 Ratification

Defining participants in conversation based on the mere co-presence, Goffman makes a distinction between *ratified* (or official) and *unratified* (or unofficial) participants. Ratified participants are those participants who “have declared themselves officially open to one another for purposes of spoken communication and guarantee together to maintain a flow of word.” (Goffman, 1967, p. 34). Therefore, the process of ratification is not conducted at the level of individual actions but at the level of collaborative actions of all participants in conversation. Furthermore, participants are ratified not to spoken interaction as ongoing process but to *a state of talk* as a naturally bounded unit.

Unratified participants are treated as someone not formally participating in conversation. Examples include spectators during a trial or an audience of a TV talk show. Although they are not intended to contribute to the discussion, in some situations they may provide some information, verbal or non-verbal feedback. Therefore, it is sometimes very difficult to maintain the border between official and unofficial participants. There is also a large disagreement among researchers regarding this issue. Kerbrat-Orecchioni (2004) illustrates this in the example of ratification of spectators and juries of a trial. Some investigators consider spectators and juries of a trial as unratified participants as they are forbidden to speak. Others, however, consider jury as very loosely speaking ratified participants and spectators as unratified participants. Kerbrat-Orecchioni (2004) claims that spectators can be also treated as a ratified participant to a certain extent given the legitimacy of their presence, their displayed interest in the trial, and the fact that they are also intended recipients of what is being said. For that reason, Kerbrat-Orecchioni (2004) introduces different:

- **degrees of ratification** - witnesses are 'more ratified' than jury who are nevertheless more ratified than spectators.
- **modes of ratification** - complete ratification in both production and reception formats vs. ratification in reception only.
- **levels of ratification**- *global level of ratification* that is defined based on the script as, for example, in a case of some institutional conversations (e.g. a city council meeting or a court trial) vs. *local level of ratification* related to a particular episode or task.

As a state of talk, in Goffman's terminology, consists of "the total activity that occurs during the time that a given set of participants have accredited one another for talk and maintain a single focus of attention " (1967, p.35), Goffman's notion of ratification defined at the level of a state of talk may refer to both global and local levels of ratification.

The process of hearing what a speaker says should be distinguished from the official status as a ratified participant in interaction (Goffman, 1981a). One might not be listening though having the status of ratified participant in interaction. Similarly, one might be following the talk closely while being an unofficial participant in interaction.

2.6.3 Categories of participation

Criticizing traditional approaches to analysis of conversational interaction based on dyadic models, Goffman emphasized the insufficiency and simplicity of the traditional notions of speaker and hearer and suggested that these terms should be decomposed into a set of categories that provide a better insight into the structure of interaction. Though the proposed categorization has some disadvantages, it has been employed as a starting point for most of the theoretical and practical research into analysis and modeling of multi-party interaction in various domains.

In this section, we first present Goffman's categorization of participation in conversation. Then, we describe several other existing taxonomies comparing them with Goffman's categorization. Finally, we present the categorization of participation roles that is used in our study.

Goffman's categorization of participation

Goffman (1981a) distinguished three basic kinds of hearers to talk: *addressed recipients*, *unaddressed recipients* and *bystanders*. Addressed and unaddressed recipients are ratified participants in interaction that are differentiated based on whether or not they are being specifically addressed by the speaker. Bystanders, on the other hand, are unofficial participants in interaction whose access to the conversation, however minimal, is perceivable by the official participants. As Goffman has emphasized, their presence should be considered the rule, not the exception. Bystanders can follow the talk in two different ways. In some situations, they can briefly follow the talk, catching some pieces of it, without much effort or intent and thus becoming *overhearers*. In other circumstances, they can use the accessability they have to listen to the conversation carefully and thus becoming *eavesdroppers*, not dissimilar to those who secretly follow the conversation electronically.

When considering the notion of speaker, Goffman (1981a) suggested a decomposition into three categories. First, the animator, someone involved in the role of utterance production. Second, the author, "the agent who puts together, composes and scripts the lines that are uttered" (Goffman, 1981c, p. 226). Third, the principal, someone who is committed to what has been said, someone whose beliefs have been told and whose position has been established. The term speaker mostly implies to the case where all three categories are used together. Nevertheless, the animator can recite someone else's text or speak *for* someone else and *in* someone else's words (Goffman, 1981a, p. 145). Furthermore, speaking participants can switch between these three roles during the course of utterance production.

It is to be noted that animator and hearer are parts of the same level of analysis. They are "not social roles in the full sense as much as functional nodes in a communication system"(Goffman, 1981a, p. 144). Therefore, the animator and three kinds of hearers represent *participation statuses* that an individual can have relative to the utterance being produced.

Animator, author and principal, taken together, comprise the *production format* of an utterance. Another basic element in the structure of an utterance is the *participation framework*. It represents the collection of participation statuses of all participants in interaction, one of which is that of animator. Goffman's categorization of participation (participation statuses and production formats) as well as its relation to the participation framework are presented in Figure 2.1.

Some investigators, however, consider the production format and the participation framework as parts of the same level of analysis linking the production format to the

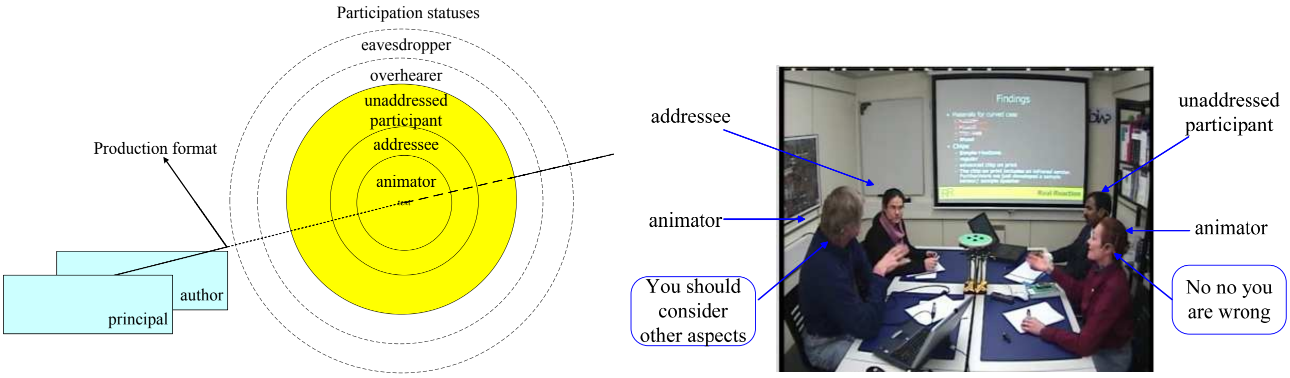


Figure 2.1: (a)Participation statuses and production formats (b) Participation framework

notion of speaker and participation framework to the notion of hearer only (Levinson, 1987; Kerbrat-Orecchioni, 2004). Considering the animator as a part of the participation framework allows that the party who is uttering words is also at the receiving end of the utterance.

Levinson (1987) pointed out several disadvantages of Goffman's categorization. First, the proposed categories are empirically inadequate as they do not provide sufficient distinctions: Barbara's status related to Ann's utterance in the example below, remains undesignated in the Goffman's categorization system. Clearly, in telling Charles that Barbara knows what's playing at the theater, Ann indirectly asks Barbara to answer Charles' question.

< 6 > (*from Clark and Carlson (1992): p. 211*)

1 Charles to Ann: Ann, what's playing at the theatre next week?
2 Ann to Charles: Sorry, I don't know. But Barbara does.
3 Barbara to Charles: "Much Ado About Nothing".

Second, the categories are not sufficiently characterized in a way that renders their applications feasible. Third, the proposed categorization is considered relevant only for ordinary conversation whereas for other forms of talk-in-interaction other categorizations are required. Levinson (1987) explains this with Goffman's failure to make a crucial distinction between application of these terms at the level of utterance event and at the level of speech event. Speech event, in Levinson terminology, refers to a unit of a specific form of talk in interaction as a whole (e.g. meetings or conferences).

Other categorizations of participation

Clark (1996) proposed categorization of participation roles similar to Goffman's categorization. This taxonomy can be considered as a modified version of categorization of participation roles proposed in (Clark and Carlson, 1992). According to Clark (1996), all people around an action can be divided into those who truly participate in the action and those who do not: participants and nonparticipants. Participants in conversation encompass speaker, addressees as well as others taking part in conversation but currently not being addressed, so called *side-participants.* All other listeners to conversations are called *overhearers.* There are two basic kind of overheares: bystanders and eavesdroppers. Bystanders are those who are openly present and the speaker is aware of their presence but they are not part of the conversation. Eavesdroppers are those who are listening to conversation without the speaker's awareness.

Clark's and Goffman's taxonomies differ mainly in two ways. First, Clark's taxonomy is applicable to other forms of conversations (e.g. phone conversation) as participation is not based on mere copresence. Second, the notions of overhearers, bystanders and eavesdroppers are treated differently. In Clark's taxonomy, all listeners that are not a part of conversation are identified as overhearers. They are differentiated as bystanders and eavesdroppers based on the speaker's awareness of their presence and involvement in the process of auditing. In Goffman's taxonomy, all unofficial participants in conversation are designated as bystanders. Their presence is considered as the rule. They are differentiated as overhearers and eavesdroppers based on the way they are following the talk.

Levinson (1987) suggested further classification of both production format and recipient roles by introducing further categories with finer distinctions. He proposed two ways of developing a systematic set of relevant categories. One way is to define a basic set of primitive categories and then to derive other participation roles in terms of basic ones. A simple schema of this kind proposed by Levinson is presented in Table 2.1. Another way to develop a systematic set of empirically adequate categories is to decompose the basic categories into terms of defining features and then use them to define more complex categories. Levinson's proposal for this kind of categorization regarding recipient roles in shown in Table 2.2.

PARTICIPANT as an underlying dimension of the categorization system has already been discussed in Section 2.6. CHANNEL-LINKAGE is considered with ability to receive a message. The feature of RECIPIENTSHIP informally refers to "someone whom a message is for; it is thus perhaps essentially a role defined by the pertinence of the *informational (or attitudinal) content*" (p. 178). Recipients are, therefore, designated destinations of the messages. Levinson emphasized that the ways in which recipients as well as addressees are distinguished from other participants are "both subtle and hardly understood". The notion of ADDRESS is also presented as puzzling. Levinson listed a set of explicit and tacit features of ADDRESS that are used to distinguish addressees from other participants. We will provide the detailed analysis of the notion of addressee, and thus the feature of address as well, in Section 2.6.5.

Categorization of participation in our study

For better understanding of the notion of addressee, it seems preferable to make a clear distinction between addressees and targets as two forms of destination of a message. The notion of target in our study excludes those persons that are intended recipients of speaker utterances but are not present in the conversational situation - ultimate destination in Table 2.2. Therefore, the notion of target encompasses indirect target and targeted overhearer

Basic categories

source = informational/illocutionary origin of message
target = informational/illocutionary destination of message
speaker = utterer
addressee = proximate destination
participant = a party with a ratified channel-link to other parties

Derived categories

producers = sources or speakers
recipients = addresses or targets
author = source and speaker
relayer = speaker who is not the source
goal = an addressee who is the target
intermediary = an addressee who is not the target
audience = participants who are not producers nor recipients
etc.

Table 2.1: Basic and derived categories of participation - Levinson(1987)

Reception roles	ADDRESS	RECIPIENT	PARTICIPANT	CHANNEL-LINK
(a)Participant roles				
interlocutor	+	+	+	+
indirect target	-	+	+	+
intermediary	+	-	+	+
audience	-	-	+	+
(b)Non-participant roles				
overhearer	-	-	-	+
targeted overhearer	-	+	-	+
ultimate destination	-	+	-	-
Useful subordinate classes:				
recipient = +RECIPIENT				
addressee = +ADDRESS				
participant = +PARTICIPANT				
hearers= +CHANNEL-LINK				
etc.				

Table 2.2: Reception roles - Levinson(1987)

(see Table 2.2).

Introducing target as a participation role, provides also for the extension of the traditional notion of indirect speech acts in a sense that they can be performed by means of direct ones not only at the same, but also at different, hearers. In the latter case the intended recipient of the direct act is marked as the addressee whereas the intended recipient of an indirect act is marked as the target. Clark and Carlson (1992) used the terms *linear* and *lateral* addressees to refer to addressees of indirect acts that are directed to the same and different hearers, respectively, as direct illocutionary acts.

As listeners' roles are the main focus of our study, we do not decompose the notion of speaker into different production roles. In summary, our categorization is based on the following distinctions:

- speaker
- addressee
- side-participant
- bystander: overhearer or eavesdropper
- target

2.6.4 Structural instability of the participation framework

Constant changes in production formats and participation statuses can be observed throughout interaction. However, interactions involving more than three official participants and/or bystanders make changes in the *framing of participation* possible in several ways. These changes result in what Goffman (1981a), calls "structural instability" of the participation framework.

A complete change can be observed when a conversation is split off into two or more conversations as well as when several conversations are merged together. This kind of instability is often found in interactions involving a large number of participants. These conversations may be dependent as one person can participate in several of them. In some circumstances, two differently managed conversations may occur under conditions of mutual accessability each bystanding the other. The right to leave and to join a conversation implies situations in which participants may shift from one conversation to another (Goffman, 1981a, p.135).

Partial changes in framing are observed when *subordinate communication* occurs in the process of *dominate communication.* Goffman (1981a) distinguish three forms of subordinate communication:

- **byplay**- subordinate communication between ratified participants
- **crossplay**-communication between ratified participants and bystanders across the boundaries of dominate conversation

- **sideplay**- communication among bystanders

Those who take part in subordinate communication may attempt to conceal that they are communicating. Goffman refers to this kind of subordinate communication as 'collusion'.

2.6.5 Addressee

The notion of addressee as well as relations among addressed and unaddressed participants are presented in the literature as complicated, puzzling and not much explored (Goffman, 1981a; Levinson, 1987; Clark and Carlson, 1992).

Goffman (1981d, p.9-10) defines addressees as those ratified participants "oriented by the speaker in a manner that suggest that his words are particulary for them, and that some answer is therefore anticipated from them more so than from the other ratified participants". According to this, the addressee is the one who is expected by the speaker to react on what the speaker says by for example, providing verbal or nonverbal feedback, by taking the turn and providing a reply or by performing a required nonverbal action. In the speaker orientation towards the addressee, Goffman insists in particular on visual cues referring to the addressee as to "the one to whom the speaker addresses his visual attention" (Goffman, 1981a, p.133).

There are two main drawbacks to Goffman's definition. First, Goffman fails to make a distinction between the feature of address and the feature of recipientship (see Section 2.6.3) and in line with that between addressee and target of a speaker's message. In the case when the addressee and the target of the message are two different ratified participants, the speaker sometimes expects that the target, and not the addressee, reacts on what he says. Second, gaze directional cues are just one means of addressing: a speaker may address someone who is not within the focus of his visual attention. Furthermore, looking at someone does not always imply addressing the gazed person. This is due to context-sensitivity of gaze-directional addressing which will be discussed in Section 2.7.2.

Similar to Goffman, Clark and Carlson (1992, p.220) identify the notion of addressee with the notion of targets, defining addressees as "ostensible targets of what is being said. Ordinarily, they are the participants for whom the speaker has the most direct and obvious goals in designing his utterance".

Following Levinson's (1987) approach in defining participation roles in terms of defining features, we distinguish the following notions: "*the message is directed to X*" and "*the message is intended for X*". A speaker may employ a variety of mechanisms to show explicitly or implicitly to whom he is directing his speech. These mechanisms can be of three kinds: verbal (language use and language structure), nonverbal (e.g. gaze and gestures) or contextual (e.g. sequential organization of talk or a state of the knowledge of co-participants). They will be discussed into more details in Section 2.7.2. Directing his speech to X, a speaker is actually giving primary attention to X. Addressees are, therefore, defined as those participants to whom the speaker is directing his speech. In a large number of cases, participants to whom a message is directed are also intended recipients of the message. However, addressed participants can be someone who is not the intended receiver of the message. For example, knowing that Ann is the only one responsible for cooking in the house and that her husband John has no experience with cooking, Mike's remark addressed to John: "Fresh basil would be a perfect additional ingredient to this meal" is not intended for John but for Ann. Similarly, Mike's question addressed to John: "Is this prepared with ginger?" is intended for Ann who is currently chatting with Mike's wife. In the case that Mike's intention is that John passes the question to Ann, John is considered as *intermediary* for the ultimate destination of the message - Ann (see Table 2.2).

In the case when a speaker's message is not directed to anyone in particular (i.e. it is broadcast), those official participants in conversations for whom the message is intended are considered as addressees. Speaker may also broadcast his message to others without knowing which of them he is addressing. Consider an utterance of the sort "The last of you, please close the door": though the speaker explicitly direct his utterance to the last one to leave the room by means of vocatives, he is treating all participants as equally potential addressees. Clark and Carlson (1992, p.235) term this kind of process as "addressing by attribution".

2.7 Recipient design and role assignment

Having defined a set of categories for participation roles, the main problem for analysts arise as to how to assign these categories to participants during the course of an utterance. This may pose problems not only for analysts but also for people who are themselves involved in interaction (Levinson, 1987; Kerbrat-Orecchioni, 2004). Detailed explorations and understanding of mechanisms that speakers use to designate which hearers are to take which roles is therefore required.

The *role assignment* is a critical part of one of the most general principles organizing talk within conversation - *recipient design*. According to Sacks et al. (1974, p.727), the recipient design represents "a multitude of respects in which the talk by a party in a conversation is constructed or designed in ways which display an orientation and sensitivity to the particular other(s) who are co-participants". Clark and Carlson (1992) employ the term *audience design* to extend the notion of recipient design to include overhearers in addition to official participants in conversation. Speakers design their utterances with different types of hearers in mind though dealing with them in different ways. Therefore, Clark and Carlson (1992) divide audience design into participants design, addressee design and overhearers design. The notion of participants, in Clark and Carlson's terminology, encompasses addressees and side-participants. In designing their utterances, speakers focus their attention primarily on addressees. Participants are intended to be informed about the act the speaker is performing towards addressees and *to recognize the meaning of that act* (Clark and Schaefer, 1992, p. 260). Overhearers are generally not meant to understand how an utterance has been designed for them. In dealing with overhearers, speakers may choose among a range of attitudes such as indifference, disclosure, concealment and disguisement (Clark and Schaefer, 1992). Levinson (1987) also investigated the recipient design principle for indirectly targeted utterance focusing on targeted participants who are not being addressed.

In this study, we are concerned only with *addressee design* - primarily, with understanding of mechanisms that speakers use to show "orientation and sensitivity" towards their addressees. Some major addressing mechanisms will be listed in Section 2.7.2 and elaborated in Section 2.10 regarding addressing in meetings.

2.7.1 Forms of addressing

Addressing represents a form of orientation and directionality of the act the current speaker performs toward the particular other(s) who are involved in interaction.

Two forms of addressing can be distinguished: *explicit addressing* and *tacit addressing.* Explicit addressing includes, for example, the use of address terms such as names or titles, where the speaker unambiguously identifies the addressee of his utterance. Tacit addressing is a manner of addressing which recipients draw upon diverse features of the content and context of the action performed by the speaker (Lerner, 2003). For instance, sequential organization of talk can provide the circumstances for accomplishing tacit addressing as it is illustrated in the following example:

< 7 > (*participants A, B, C*)
1 A to B Have you ever been there?
2 B to A No
3 ⟹ A to B Why?

As the question at line 3 represents a follow-up question to the question at line 1, it can be seen as tacitly addressed to the same participant (B). Similarly, in the conversation involving three participants, a speaker can tacitly address one participant by referring to another participant by name and thus excluding him as a possible addressed participant: "Did you know that Ann is very good in cooking?".

2.7.2 Addressing behavior

Addressing behavior is behavior that speakers show to express to whom they are addressing their speech. It depends on the course of the conversation, the status of attention of participants, their current involvement in the discussion as well as on what the participants know about each other's state of knowledge, whether explicit addressing behavior is called for. In this section we present various aspects of addressing behavior that are relevant for face-to-face conversations. However, in Section 2.10 we will discuss how different addressing practices are affected by the meeting context.

In face-to-face conversation, addressing is carried out through various communication channels such as speech, gaze and gestures. As addressing denotes a form of orientation and directionality of the act the speaker performs toward particular others, those aspects of verbal and nonverbal behavior that bear deictic reference to the participants present may be considered as explicit addressing mechanisms. However, as we will discuss later, these mechanisms may also serve different purposes in conversation and may be context sensitive. Moreover, the success of nonverbal addressing behavior of a speaker is contingent on the behavior and the attention of his co-participants.

As previously discussed, speakers may accomplish addressing tacitly by means of sequential organization operating in conversation. In many cases, the person addressed will be the person who last spoke or even more so, the person whose speech was the stimulus for the present response. Participants' physical arrangement and individual or group nonverbal activities in which they are involved during the conversation can also be added to the shared particularities of circumstances and context that are employed in the process of tacit addressing.

The state of a participant's knowledge about the reported event can also play an important role in the addressee design. Analyzing conversations where both knowing and unknowing participants were co-present, Goodwin (1981) found that an unknowing participant, the one who lacked relevant information about the event, was addressed while the knowing recipient was asked to monitor what the speaker was saying for its correctness. However, when the speaker moved his gaze towards the knowing participant, he displayed uncertainty about what he was saying by producing a request for verification appropriate for a knowing recipient.

In the following sections, we present in more detail verbal and non verbal addressing practices.

Verbal aspects of addressing behavior

Addressing can be accomplished partially or completely by using resources obtained from an utterance's content. In the case of partial addressing, the content indicates that someone is being addressed without revealing who that participant is. Furthermore, this manner of addressing is not limited only to utterances addressed to an individual; it can also be employed for addressing a group of participants. In addition to the utterance content, a type of activity the speaker is performing with his speech can be an additional resource for distinguishing among addressing types (see Section 2.8): some types of moves - such as agreements or disagreements - tend to be addressed to an individual rather than to a group. More information about the addressee of a move can be induced by combining the information about the type of the move with so-called "lexical markers" within the move that are denoted as addressee "indicators". In this section, we discuss several addressing practices which make use of utterance content.

Address terms The strongest method of addressing available is the use of address terms. They encompass the usage of personal names, terms of endearment (e.g. 'darling') or categorical terms (e.g. 'boss', 'mother') in the vocative form (Lerner, 2003). It is to be noted that these categories when not used as address terms are employed for refereing to a third person, that is to a person other than the speaker and the addressee or to a person who belongs to the addressed group though not being explicitly addressed, as illustrated in the following example:

< 8 > (*participants Ann, John, Mike*)
1 Ann to John — Mike can help you with that
2 Ann to John and Mike — We should discuss John's proposal
3 Ann to John — Send my regards to Tom

Speakers employ address terms when they want to unambiguously direct their speech to a particular participant or to a group of participants. However, the usage of the address term is not always sufficient for designating who is being addressed. A 'situationally shared address term' (e.g. 'son' in some occasion) is an incomplete addressing method since it requires additional sources of information to be employed for determining who of the participants referred to is the addressed one (Lerner, 2003).

Although being the most explicit addressing method, address terms are not widely used. Moreover they seem to be used primarily under specific circumstances in which they are employed to do more than just simple addressing (Lerner, 2003). The additional usages of address terms are mostly manifested through different positioning of address terms within a turn. For example, pre-positioned terms of address in many situations are employed for verifying or establishing the availability of the recipient in situations where this may be problematic. To ask Ann, who is currently taking notes, to give him a book, John may address Ann in the following way "Ann, can you give me the book next to you". Post-positioned terms of address are often used as repairs, that is, when the success of other methods over the course of turns is doubtful. Those and other usages of pre- and post-address terms are described in more detail in (Lerner, 2003).

Linguistic markers Addressing a single participant or a group of participants often encompasses the use of personal pronouns as an example of person deixis. More precisely, the second person deixes (e.g. you, your, yours) and the first person deixes (e.g. we, us, our, ours) are employed as a means for partially accomplishing addressing. In multi-party conversation, the use of the first and second person references do not automatically resolve who is being referred to. Therefore, additional sources of information are required for resolving the ambiguity.

The second person deixis (henceforth, *you*) refers to a single participant or to a group of participants excluding the speaker. Commonly, the referred participant is the addressed one. However, *you*, may have usages other than referring to the addressee. An overview of such usages is given in (Schegloff, 1996). *You* can be used explicitly for addressed summoning (e.g. "Hey you!"). Also, *you* is sometimes used for referring to a category of persons or to everyone. The following example illustrates this kind of so-called impersonal usage of *you*:

< 9 >
1 A: People like to be recognized for the quality of their work.
2 B: Sure. A positive feedback motivates *you* to do *your* job even better.

When *you* is employed for the conjoined task of person reference and establishing addressing, *you* is said to be an *addressee indicator* but not an *addressee designator* (Lerner, 2003). Various practices for determining the referent of *you* are described in (Lerner, 1996b). The reference can be sometimes resolved from the specifics of the situation, identities, and particularities of content and context of the talk. In some instances, sequential positioning of a turn may provide sufficient information for determining who is being addressed by *you.* Visible aspects of interaction including participants' visible group or individual actions can also be employed for resolving the reference of *you.* In addition to all these features of the talk and its circumstances, who is the person referred to by *you* is in most cases determined by explicit addressing practices such as gazing at the person. As will discussed below, gaze, when used alone, can be a weak or troublesome addressing device. However, it can be seen as an 'enhanced' addressing method when combined with *you* (Lerner, 1996b). As presented in (Jovanovic and op den Akker, 2004), additional linguistic resources such as determiners, numerals and indefinite pronouns can be employed for distinguishing whether *you* refers to a single participant or to a group of them (e.g. two of you, some of you, all of you or anyone of you).

A speaker uses the first person deixis *I* (and its variants such as *me* or *mine*) for self-reference and *we* (and its variants *us* or *our*) for referring to a group of people including the speaker. The latter is of particular importance for determining who is being addressed by the speaker. There are two main issues concerning the recognition of the referents of *we* other than the speaker: the size and members of the group referred to. The group referred to with *we* may include: (a) the speaker and the addressee(s), (b) the speaker, the addressee(s) and other persons who may or may not be participants in the conversation and (c) the speaker and non-addressed persons who may or may not be participants in the conversation. The following example illustrates all three types of usages of *we*:

< 10 > *participants Ann, John and Mike*

1	a	Ann to John	*We* can visit Mike tomorrow.
2	a	Ann to John and Mike	Shall *we* meet in one hour?
3	b	Ann to John	*We* have to go with your car, as Mike's car needs to be repaired.
4	b	Ann to John and Mike	Tom will arrive tomorrow. So, *we* can together go out for a dinner
5	c	Ann to John and Mike	Yesterday I saw an old friend of mine. *We* haven't seen each other for years.
6	c	Ann to John	I had discussed with Mike and *we* agreed not to go to the cinema.

Features of the talk and its circumstances similar to those used for determining the referent(s) of *you* listed above, may be relevant for identifying whether the group referred to with *we* includes participants in conversation and even more so, the addressed participant(s). Furthermore, the type of the act performed in speaking can be useful in distinguishing whether or not the participants are referred to. For example, when *we* is used in combination with suggestions or in combination with elicitation of any kind of verbal or non-verbal action, *we* refers to a group containing co-participants (see (1), (2), (3) and (5) in the example 10).

Nonverbal aspects of addressing behavior

Gaze Analyzing dyadic conversations, researchers into social interaction observed that gaze in social interaction is used for several purposes: to control communication (e.g. turn-taking), to provide a visual feedback, to communicate emotions and to communicate the nature of relationships among interacting participants (Kendon, 1967; Argyle, 1973).

Recent studies into multi-party interaction emphasized the relevance of gaze as a means of addressing. Vertegaal (1998) investigated to what extent the focus of visual attention might function as an indicator for the focus of "dialogic attention" in four-participants face-to-face conversations. "Dialogic attention" refers to attention while listening to a person as well as attention while talking to one or more persons. Empirical findings have shown that when a speaker is addressing an individual, there is 77% chance that the gazed person is addressed. When addressing a triad, speaker gaze seems to be evenly distributed over the listeners in the situation where participants are seated around a table. It is also shown that on average a speaker spends significantly more time gazing at an individual when addressing the whole group, than at others when addressing a single individual. When addressing an individual, people gaze 1.6 times more while listening (62%) than while speaking (40%). When addressing a triad the amount of speaker gaze increases significantly to 59%. According to all these estimates, it is to be expected that gaze directional cues are good indicators for addressee identification in face-to-face conversation.

However, these findings cannot be generalized in the situations where some objects of interest are present in the conversational environment, since it is expected that the amount of time spent looking at the persons will decrease significantly. As shown in (Bakx et al., 2003), in a situation where a user interacts with a multimodal information system and in the meantime talks to another person, the user looks most of the time at the system, both when talking to the system (94%) and when talking to his co-participant(57%). Also, another person looks at the system in 60% of cases when talking to the user.

A physical or seating arrangement of interacting participants is implicated in organization of gaze as a means of addressing. Moreover, it specifies a so-called visible area of each participant. During a turn, a speaker mostly looks at the participants who are in his visible area. Moreover, the speaker frequently looks at a single participant in his visual area when addressing a group. However, when he wants to address a single participant outside his visual area, he will often turn his body and head toward that participant.

Lerner (2003) has investigated the context sensitivity of gaze as an explicit addressing device when combined with a speaker initiating action, that is, with a first pair part of an adjacency pair for the selection of the next speaker. It has been shown that a gaze as an addressing device is vulnerable to looking practices of recipients: (1) it may not be seen by the addressed recipient and (2) it may not be seen by a side participant although the speaker and the addressee have established mutual gaze which in turn may cause that the side participant speaks next.

Gestures Deictic hand and head gestures are usually employed by a speaker to explicitly demonstrate to co-participants that his utterance is being directed to a particular participant. However, the addressee of a speaker's utterance is not necessarily the same person as the person whom the speaker points at.

Commonly, non-verbal deictic devices are accompanied with a verbal reference to the person the speaker is pointing at, that is, with a person deixis. Examples include the second and third person deixis, the name or the title of the person. When a deictic gesture is accompanied by a second person deixis or by an addressed term referring to the target of the gesture, the targeted person is the addressee of the speaker utterance. Furthermore, the deictic gesture is accompanied with gazing at the same target. In all other cases, deictic gestures are employed for referring to a person without explicitly addressing that person: the person pointed at is the one the speaker is talking about having the role of a side participant or of a target or belonging to the addressed group though not being explicitly addressed. Furthermore, when talking about the person a speaker is pointing at, it is not always the case that the speaker is gazing at that person. The following example illustrate the discussed usages of hand and head deictic gestures:

< 11 >

1 John to Ann: Is this your [*pointing at Ann, gazing at Ann*] book?
2 Ann to John: No, it is his [*pointing at Mike, gazing at Mike*].
3 Mike to John: I suggest that Ann
[*pointing at Ann, gazing at John*]
read first my book and then give it to you.
4 John to Ann and Mike: Okay, then we agreed first she
[*pointing at Ann, gazing at Ann*] and then me.

Non verbal aspect of speech Addressees can also be designated by the manner of speaking (Clark and Carlson, 1992). For example, by whispering, a speaker can select a single participant or a group of participants as addressees. Similarly, by raising his voice, a speaker can denote that he is addressing a particular participant or a group of participants who are not in close proximity.

2.8 Addressing types

A speaker may address his speech to the whole group of official participants, to a subgroup of them or to a single participant in particular in an explicit or tacit manner. In that regard, we distinguish three types of addressing: *individual addressing*, *subgroup addressing* and *group addressing*. As will be presented in Chapter 4, observers of meeting conversations had particular difficulties in distinguishing subgroup from group addressing indicating that the concept of subgroup addressing is rather vague.

Utterances addressed to more than one participant can be further classified as being *distributively* and *collectively* addressed. The question "Do you have children?" addressed to a couple is an example of collective addressing whereas the same question addressed to participants who are not married is considered as distributed among participants. A response to a collectively addressed question is expected from any of the addressed participants whereas a response to a distributively addressed question is expected from each participant being addressed. Lerner (1993) has extensively investigated various types of collectivities that can be constituted in conversation as well as different practices for establishing relevance for conjoined participation. He uses the term "association" to refer to any assembly of co-present individuals that are in any of various ways classified as a collectivity.

It is also possible that the incumbent of speaker and addressee roles is the same person. Uttering softly "What else do I want to say?" while trying to evoke more details about the issue that he is presenting, a speaker is talking to himself without really addressing other than himself. Furthermore, other participants are not considered as recipients of the message as it was not formulated *for* them. This kind of addressing is refereed to as *self-addressing*.

Levinson (1987) pointed out that there are certain types of utterances that do not seem to presuppose any role other than speaker. Certain expressives such as "Oops!", "Ouch!", "Eek!" do not seem to be addressed to self or others though they may have indirect targets. Levinson (1987) calls these vocalization 'out-louds' whereas Goffman (1981e) uses the term 'response cries'. According to Goffman (1981a), when uttering these kinds of 'response cries', one may intend to provide information to everyone in the range seeking some response from them, but not a specific reply. The intended recipients in these situations are not hearers but overhearers - targeted overhearers in Levinson's terminology (see Table 2.2). Further examples of utterances that do not require co-present addressee or side-participants are conventional blessings or curses (Levinson, 1987).

In our study on addressing presented in this thesis, we do not consider the distinction between collectively and distributively addressed utterances. The study also excludes self-addressing and 'out-louds'.

2.9 Relations among recipient design, turn-taking and adjacency pairs

As a speaker designs his utterance with different types of hearers in mind, he actually performs different doings to parties present. Therefore, a multi-party situation is seen as one "where we are dealing with the relationship of activities - not sequentially [...]- but across the parties present, given an activity done directed to some other person" (Sacks, 1992, p.530). In this section, we are concerned with analysis of relations among turn-taking organization, sequential organization of actions and organization of actions across parties.

As discussed in section 2.3.1, one of the 'current-selects-next' techniques is the use of the first-pair part of an adjacency pair that employs certain addressing practices. First-pair parts contribute to current speaker's techniques for selecting next speaker as they set constraints on what should be done in a next turn. They, however, do not allocate by themselves a next turn to any specific participant. Also, addressing someone does not necessarily imply that the addressed participant is selected to speak next. For example, a speaker may address his answer to the participant who asked a question without selecting him to speak next. Therefore, only the use of first-pair parts in combination with addressing practices leads to the selection of the next speaker (Sacks et al., 1974).

The current-selects-next techniques are workable only in situations when addressing practices limit eligible responder to a single participant. As previously discussed, the turn-taking model does not provide for the selection of the next speaker when subgroup and group addressing practices are employed. Furthermore, utterances directed to a particular individual can be designed in a way that indicates that some other party is indirectly targeted. In some cases, indirect targets seem to be the immediate responder and not the one who was addressed. Thus, when first-pair part is designed in this way, targeted participants and not the selected one may speak next.

The SSJ turn-taking model has been designed as being *context-free* but capable of *context-sensitivity*. Recipient design, as the most general principle that particularizes conversation, is a place where "aspects of situatedness, identities and particularities of content and context" (Sacks et al., 1974, p.699) can shape 'current-selects-next' allocation techniques. Lerner (2003) described a context-sensitive operation of addressing practices employed by the current speaker to select the next speaker. Regarding explicit addressing, Lerner emphasized context-specific limitations of gaze-directional addressing used in combination with first-pair parts in the selection of the next speaker (see Section 2.7.2).

However, relations among recipient design, turn-taking and adjacency pairs are best described through the accomplishment of tacit addressing and thereby tacit selection of next speaker. Sometimes an utterance that is designed as a first-pair part can be understood as being addressed to a single recipient based on the design and content of the utterance or the circumstances relevant for its production such as identities, situation, course of actions or sequential position. This form of recipient design tacitly accomplishes addressing and thereby contributes to the tacit selection of the next speaker. On the other hand, sequential organization of talk and turn-taking design that locally operate on the current and next turn can provide circumstances for accomplishing tacit addressing. Lerner (2003) illustrates this with 'next-turn repair initiators' such as "*What?*" or "*Hm?*". These kind of first-pair parts detect a source of a problem in the prior turn and makes a repair of that problem relevant for next turn. In locating the problem in the prior turn, 'next-turn repair initiators' address the problem in the production of the prior turn, addressing in this way the producer of that turn. In these kinds of situations, the turn-taking features such as 'just prior turn' and 'just prior speaker' provide the context for accomplishing addressing tacitly (Lerner, 2003).

So far, we have been focused only on the first-pair parts of adjacency pairs and their contributions to the selection of the next speaker. However, adjacency pairs by themselves may provide the relevant context for accomplishing tacit addressing. In many cases, the second-pair parts are addressed to the speaker of the first-pair part: an answer addressed to the person who asked a question, an acceptance addressed to the person who offered to perform an action. However, it does not imply that the speaker of the first-pair part is always the addressee of the second-pair part. For example, A can address a question to B whereas B's reply to A's question can be addressed to the whole group. In this case, the speaker of the first-pair part is only one of the incumbents of the addressee role of the second-pair part. There may also be cases, when the speaker of the first-pair part is not addressed in the second-pair part, as in the following examples:

< 12 >

John to Ann and Mike: We might want to have the TV remote control in red.
Ann to John: Yes, that's a very good idea.
⟹ Mike to Ann: I don't agree with you, Ann.

< 13 >

John to Mike: You should prepare a presentation for tomorrow.
Ann to John: with much more details
⟹ Mike to Ann: I will do that, Ann.

Although Mike's utterances in both examples are second-pair parts of adjacency pairs with John's utterances as first-pair parts, they are both explicitly addressed to Ann. With his utterance in the first example, Mike is performing two actions across parties: in disagreeing with Ann's agreement to John's proposal, Mike is actually disagreeing with John who is thereby the intended recipient of Mike's message though not being addressed. The second example is an example of piggybacking, where Ann's utterance is built upon John's proposal. More precisely, the action that Ann performs with her utterance is an elaboration to John's proposal directed to the addressee of the proposal - Mike. Though Mike's utterance is explicitly addressed to Ann, it is also intended for John as John's proposal and Ann's elaboration can be seen as a collaborative action performed to Mike.

2.10 Participation framework and recipient design in face-to-face meetings

As presented in Section 2.4, face-to-face meetings and conversations are two forms of talk-in-interaction that may differ across several dimensions. We have previously described differences in turn-taking organization and overall structural organization. They are mostly influenced by the formality of meetings that can vary on the scale from highly formal to informal. In general, meetings are more tightly structured than ordinary conversations, with clear purpose of gathering and with a ratified set of participants who may have been assigned different roles in meetings (Goffman, 1983). In this section, we discuss how the overall organization of meetings and varieties of tasks in which participants may be involved during meetings can affect the participation framework and the recipient design.

In meetings, participants may perform various kinds of group activities that are characterized with a single speaking slot of a participant who has been given an exclusive claim to the floor for some period of time. Examples include presentations, demonstrations, opening or closing a meeting. These kinds of *meeting activities* have some properties of those social arrangements that Goffman (1981a) called *podium events* or *stage events.* Podium events such as a platform monologue or a play are performed by orators or actors. Opposite to a speaker in conversation who has fellow conversationalists, actors and orators have an *audience.*

An audience hears a talk in a different way than conversationalists in conversations, in which organization of talk is performed on turn-by-turn bases. The role of the audience is to examine and appreciate the talk, not to reply in any direct way. Their contributions to the talk in progress are mostly limited to verbal and non-verbal feedbacks. However, in some cases, the audience or some audience members may get the floor (Goffman, 1981a). In some less formal meetings, a presenter may ask the audience member(s) a question and similarly, the audience members may interrupt the presenter by asking him a question.

In a similar way, the talk for conversationalists in conversation is designed differently than for the audience of stage events. Regarding, for example, presentations in meetings, a presenter shows a different kind of addressing behavior than a speaker during a discussion: regardless of the fact that a presenter has turned his back to a participant in the audience during a presentation, he is most probably addressing his speech to the whole audience including that participant, whereas the same behavior during a discussion, in many situations, indicates that that participant is unaddressed. Similarly, the relevance of the gaze as an aspect of addressing behavior is influenced by the meeting activity currently being performed. For example, when giving a presentation, a presenter may only look at a single participant in the audience although addressing everyone present.

Examining further the notion of audience, Goffman (1981a) distinguished the audience that hears the talk on TV or radio from the live audience i.e. live witnesses. Live witnesses are considered as addressed co-participants, not addressed conversationalists. Radio or TV talk is usually not addressed to the audience, who is considered as overhearer, but to some imagined recipients according to whom the speaker styles his talk. There are also cases where two or more actors are on the stage preforming speech by addressing each other (e.g. a play). A live audience in these situations is considered as overhearer. In meetings, the live audience consisting of ratified participants is addressed. However, there may be some unofficial participants present in the meeting environment who overhear the talk (e.g. technical support staff).

In addition to podium events, meetings are also characterized by activities that resemble an ordinary conversation such as brainstorming discussions, problem solving, decision making, or negotiations. The participation framework and the recipient design for these types of activities are similar to those of conversations.

At any point in meetings, participants may also be engaged in some sort of tasks such as finding the folder that contains presentation slides of the following presenter or distributing hand notes among participants. Words spoken in these occasions are an integrated part of the activity in progress. Verbal directives are usually answered with non-verbal activities and non-verbal activities are often answered with verbal responses; an utterance is hardly context of another utterance (Goffman, 1981a). In the following example, A is trying to find a folder that contains B's presentation slides; the organization of a folder is projected on the projector screen; B is located in front of the projector screen; C and D are sitting and looking at the screen:

< 14 >

1	B to A:	Go to the folder AMI-Meeting.
2	A:	*(opening the folder AMI-Meeting)*
3	C to A:	Okay. And now the folder Presenter1.
4	A:	*(dragging the mouse over the folder Presenter2)*
5	D to A:	a bit up
6	A:	*(opening the folder Presenter1)*
7	B to A,C,D:	Here it is finally.

This example shows that the coordinated task, not conversation, is the relevant context of words. As such, it is employed for accomplishing tacit addressing.

In summary, "the notion of a conversational encounter does not suffice in dealing with the context in which words are spoken; a social occasion involving a podium event or no speech event at all may be involved, and in any case, the whole social situation, the whole surround, must always be considered". (Goffman, 1981a, p.144)

As discussed in Section 2.6.3, Levinson (1987) criticized Goffman's approach in defining the basic categories of participation as applicable only to ordinary conversation and introducing other kinds of participation roles such as actor and audience for podium events. He explains this with Goffman's failure to make distinction between application of the basic categories at the level of utterance event and at the level of speech event. We agree to Levinson's criticism regarding the assignment of participation roles to moves performed while speaking. However, we believe that Goffman's distinction between the audience of podium events in meetings and conversationalists in conversation-like meeting activities is important from the aspect of recipient design and thus for manual (see Chapter 3) as well as automatic (see Chapter 5) assignments of participation categories to interactional moves.

2.11 Participation in conversation from conversational analysis and small group research perspectives

Conversation is characterized as (a) rule governed and (b) a venue in which participants are differentiated in how they act and they are acted toward (Gibson, 2003). The rule governed aspect of conversation is presented in previous sections regarding the operation of sequential rules that limit who can speak, what they can say and whom they can address in the next turn. However, not all participants in conversation are dealt with in the same way in terms of opportunities to speak and to be addressed. These two features of conversation, are "in tension, and the tension is the greatest at the structural level of participation" (Gibson, 2003, p.1336).

Two traditions in interaction research, conversational analysis (CA) and small group research (SGR), have addressed these two aspects of conversation separately. On the one hand, conversational analysts have explored rules that organize conversational interaction and ensure a basic level ordering (e.g. turn-taking and sequence organization). Regarding the concept of participation, conversational analysis contributed to the understanding of how talk is organized across parties. The basic object of their study is the structure of interaction *per se*. On the other hand, small group researchers are mainly focused on the functioning and structure of small groups. They exploit the quantitative patterns - for example, how frequent participants speak and are addressed over some period of time - in order to gain insights into matters such as differentiation of roles, leadership or group dynamics (Bales, 1950; Shaw, 1981; Hollander, 1985). In SGR tradition, interaction is examined as a carrier of other social phenomena.

2.11.1 P-shifts - bridging a gap between CA and SGR

Gibson (2003) proposed a framework for the analysis of interaction sequences that captures both aspects of conversations: operations of conversational rules and conversational differentiation of participants. The framework is based on the concept of **participation shift** (**P-shift**), that refers to “the moment-by-moment shuffling of individuals between participation statuses” of speaker, target and unaddressed recipients” (p. 1335). In other words, it represents moment-by-moment reshuffling of the participation framework. A complete list of P-shifts, as described in (Gibson, 2003), is given in Table 2.3. The notion of target in Gibson’s terminology refers to the notion of addressee in our categorization system. However, Gibson does not clarify whether the notion of group addressing includes any number of participants or whether it refers only to the group as a whole.

As shown in Table 2.3, Gibson (2003) classify P-shifts according to the way in which the second speaker (speaker$_2$) acquires his turn:

- turn receiving - a person takes the turn after being addressed
- turn claiming - a person takes the turn after someone else addressed the group
- turn usurping - a person takes the turn after someone else has been addressed
- turn continuing - a person who has the floor changes the addressee of his speech.

P-shifts reflect any change in the position of speaker or addressee and in line with that in the position of unaddressed participant. Moreover, they are distinguished by the precise transformation of the participation framework that each of them reflects. Furthermore, P-shifts reflect not only the actions of taking, giving and losing the turn, but also different types of actions people do in speaking and addressing and things that are done to them (Gibson, 2003, p.1343). For example, AB-BA is the structural form of adjacency pairs in which the addressee of the second part is the speaker of the first part and AB-XB is the structural form of piggybacking. However, most of P-shifts are empirically unexplored in terms of interactional patterns that each of them govern.

P-shift	Example
Turn receiving	
AB-BA	John talks to Mary, then Mary replies
AB-B0	John talks to Mary, then Mary addresses the group
AB-BY	John talks to Mary, then Mary addresses Irene
Turn claiming	
A0-X0	John talks to the group, then Frank talks to the group
A0-XA	John talks to the group, then Frank talks to John
A0-XY	John talks to the group, then Frank talks to Irene
Turn usurping	
AB-X0	John talks to Mary, then Frank talks to the group
AB-XA	John talks to Mary, then Frank talks to John
AB-XB	John talks to Mary, then Frank addresses Mary
AB-XY	John talks to Mary, then Frank addresses Irene
Turn continuing	
A0-AY	John talks to the group, then John addresses Mary
AB-A0	John talks to Mary, then John makes a remark to the group
AB-AY	John talks to Mary then to Irene

The initial speaker is always labelled A, the initial target B, unless the group is addressed (or target was ambiguous), in which case the target is labelled 0. A shift has the form [speaker$_1$][target$_1$]-[speaker$_2$][target$_2$] with A and B appearing after the hyphen only if initial speaker or target serves in one of these position after the shift. When the speaker after the shift is someone other than A or B, then X is used. When the target after the shift is someone other than A, B or group, Y is used

Table 2.3: P-shifts - Gibson (2003)

Gibson (2003) also discussed how participants can be differentiated in terms of their quantitative involvement in particular P-shifts - *P-shift roles.* A person's P-shift involvement can be summarized as a *P-shift profile* that has the form of a vector of conditional probabilities that particular P-shift will be completed upon its first part, when the individual in question appears as speaker or target at least once in P-shift. For example, $P(BA|A_iB)$ represents conditional probability that person i after having addressed someone will be immediately addressed by that person. The index i is used to denote the individual in question. For each P-shift, there are several possibilities for an individual to be involved in it. These possibilities are referred to as *P-slots.* For example, AB-XA has three P-slots: A$_i$B-XA, AB$_i$-XA and AB-X$_i$A where the index i is used to denote the first occurrence of the individual in the P-shift. Having defined P-shift profiles for each participant, P-shift roles can be identified inductively by determining how people are typically differentiated from one another in their P-shift profiles. Applying some clustering algorithms leads to the identification of similar P-shift profiles, and thus P-shift roles, along a set of relevant dimensions. As conversational differentiation is not an object of our study, we will not elaborate further on it.

The P-shift framework is a powerful means for analyzing conversational sequences. It provides better insight into the structure of conversation comparing to models for examining sequences of speaking turns in group discussions (Parker, 1988; Stasser and Taylor, 1991). To make it as a more powerful tool for exploring conversational structure, the framework can be extended to capture not only changes in the participation framework but also changes in types of moves the speaker is performing towards the addressees. As Goffman (1981e) noted, the classification of moves performed in speaking provides us with an opportunity to unfold the structure of interaction that is represented as sequences of moves. Gibson (2003) also emphasized that coding of actions might be useful to advance the understanding of what precisely P-shifts are used to do. However, using a complete set of categories that can be employed to classify moves, such as speech act types, increases the number of P-shifts considerably. Within a turn, a speaker may also perform several moves of the same or different types without changing the participation framework. Moreover, with a move, a speaker may perform different types of doings and thereby have several categories assigned to it. Therefore, the future work on the P-shift framework may include a formalization of the changes in the type of actions participants are performing with their talk and its incorporation in the defined P-shifts.

2.12 Summary

This chapter discusses the concept of participation in conversational interaction taking place in a face-to-face setting. Since a great amount of research into conversational organization is based on the two-party model, the concept of participation has been somehow put aside, although many researchers were aware of its importance.

Starting from Goffman's theory of the participation framework which provides a general but incomplete insight into the concept of participation in multi-party conversation, we tried to go a step further by (1) refining Goffman's categorization system based on the method proposed in (Levinson, 1987), (2) providing a more detailed and more operational definition of the notion of addressee, (3) providing a better understanding of how addressing actually works, that is, which addressing mechanisms speakers employed in order to designate which participants are to take the addressee role and how these mechanisms are sensitive to context. We also discussed relations among addressing practices, turn-taking design and adjacency pairs organization. As the main domain of our study on addressing behavior presented in this thesis is face-to-face meetings, we described the influence of the meeting context on all relevant aspects of participation introduced for face-to-face conversation.

The operational definition of addressee, the identified aspects of addressing behavior used for accomplishing not only explicit but also tacit addressing in face-to-face meetings and identified relations between addressing on the one hand and turn-taking and adjacency pairs on the other hand are the outcomes of the analysis presented in this chapter that are of particular importance for our research on addressee identification in face-to-face meetings presented in the following chapters. They are utilized for the design of addressee annotation schemas employed in the creation of the M4 and AMI meeting corpora (Chapter 3). The findings are also exploited for motivating the selection of the features for automatic addressee identification (Chapter 5).

Chapter 3

The M4 and AMI meeting corpora

Our study on modeling addressing behavior in multi-party interaction is based on two multi-modal meeting corpora collected in the research programs of the M4 and AMI projects.

The AMI meeting corpus was developed to support multi-disciplinary research in the AMI project (Carletta et al., 2005a). Addressing is just one of a range of research topics covered in the context of AMI. As addressee annotation as well as annotation of phenomena relevant to addressing are time-consuming and expensive tasks, only a small part of the large AMI data collection has been annotated with addressee information. Since that part of the AMI corpus has recently become available for our research on addressing, we previously developed a small multi-modal corpus of hand-annotated meetings collected mostly in the context of the M4 project - the M4 corpus. The choice for using the M4 meetings, although they are short and scripted, was influenced by the fact that the meeting corpora available at that moment - ICSI (Janin et al., 2004) and ISL (Burger and Sloane, 2004) corpora - were based on audio data only.

The M4 and AMI meeting corpora were created using the NITE XML Toolkit (NXT), an open source library that supports the creation and analysis of highly structured multi-modal language corpora (Carletta et al., 2005b). This chapter starts with a brief description of NXT (Section 3.1). To make annotated data credible for obtaining valid research results, it is important to measure reliability of annotation schemas. Theoretical considerations on how to assess reliability of annotation schemas are presented in Section 3.2.

The design of the M4 annotation schemas was motivated by analysis of the addressing practices presented in Chapter 2. Although various aspects of addressing practices have been identified, the M4 scheme contains only annotations of dialogue acts, adjacency pairs, addressees, gaze direction and meeting actions. The M4 corpus is described in detail in Section 3.3. Apart from the corpus description which includes the description of the meeting data and annotation scheme, we also report the reliability of the overall annotation scheme.

The AMI corpus was developed to enhance research in various areas including speech recognition, computer vision, discourse and dialogue modeling, content abstraction, human-

human and human-computer interaction modeling. It contains a range of annotations including, among others, speech transcription, dialogue acts, addressee, topic segmentation, focus of attention, individual actions and summaries. The AMI corpus is presented in Section 3.4. After providing a short overview of the entire AMI corpus, we focus only on the description of the parts of the AMI corpus that have been employed for our research on automatic addressee classification in meetings. This also includes detailed depictions of those schemas that were used for annotation of the features employed in our classification models that will be presented in Chapter 5. In addition, we overview the reliability analysis of the presented schemas. A comparison between the AMI and M4 corpora is presented in Section 3.5.

3.1 NXT

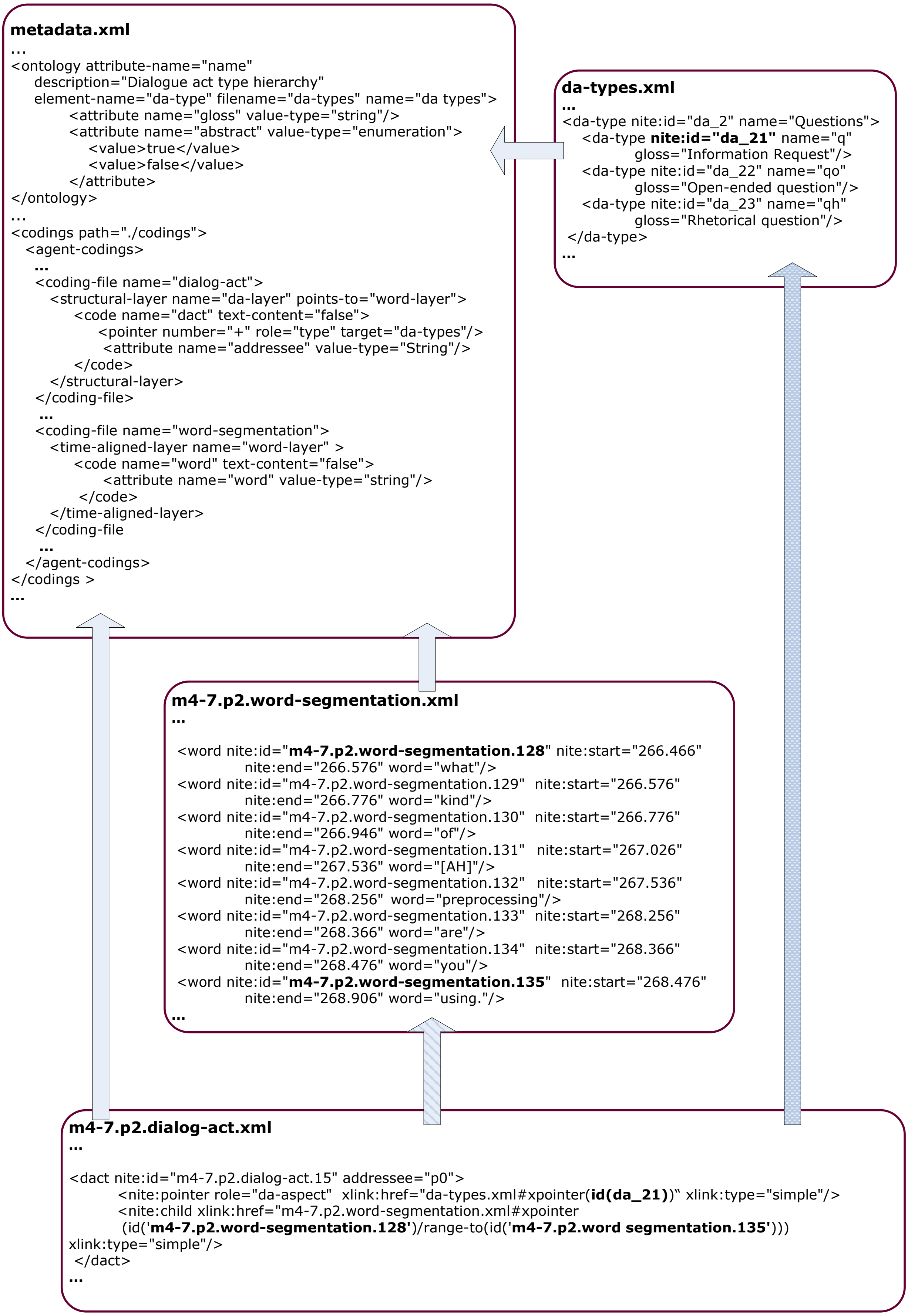
NXT is an open source software implemented in Java that provides library support for the creation, display and analysis of highly structured and cross-annotated multi-modal language corpora.

NXT supports the representation of (1) time-aligned annotated data, such as words or gaze elements, directly related to signals and (2) hierarchically structured annotated data organized using different types of relationships. Annotated data can be organized into hierarchies by structural dominance, that is, using the parent-child relationship or into more complex structures using pointers with specified roles (Carletta et al., 2005b). In the latter case, annotations may point not only to other annotations but also to entities that specify, for example, entries in a dialogue act type ontology or objects in the meeting room at which participants can look. These entities are also modelled in NXT. Furthermore, NXT provides for intersections of hierarchies as an element can have as parents elements from several distinct hierarchies. For example, words that make up an orthographic transcription may have as parents both dialogue act and named entity elements. Annotations that are built on top of other annotations may have an implicit timing information that is derived from time-aligned annotations on which they are based. Timing information is coherent across the data set.

NXT stores data in a stand-off XML data format which is represented as a set of interrelated XML files. The structures and the locations of the files as well as the relationships between them are specified in a "metadata" file. The metadata file, which formally describes a corpus, is based on the NXT Data Set Model. For the detail description of the data model, we refer to (Carletta et al., 2005b).

To gain an insight into the way a corpus is specified according to the model and the way data are stored in NXT, we provide in Figure 3.1 an illustration of a segment of the metadata file that describes the M4 corpus as well as some segments of the corresponding XML files in which data are stored[1]. The metadata file presented in Figure 3.1 shows that annotations ("dact" and "word") are organized into layers ("da-layer" and "word-layer").

[1] For clarity and simplicity purposes, we provide a modified version of the original M4 metadata file; the original metadata file specifies word-layer to point to asr-layer



metadata.xml

```
...
<ontology attribute-name="name"
     description="Dialogue act type hierarchy"
     element-name="da-type" filename="da-types" name="da types">
            <attribute name="gloss" value-type="string"/>
            <attribute name="abstract" value-type="enumeration">
                <value>true</value>
                <value>false</value>
            </attribute>
</ontology>
...
<codings path="./codings">
   <agent-codings>
     ...
     <coding-file name="dialog-act">
        <structural-layer name="da-layer" points-to="word-layer">
             <code name="dact" text-content="false">
                  <pointer number="+" role="type" target="da-types"/>
                   <attribute name="addressee" value-type="String"/>
             </code>
        </structural-layer>
     </coding-file>
      ...
     <coding-file name="word-segmentation">
        <time-aligned-layer name="word-layer" >
             <code name="word" text-content="false">
                   <attribute name="word" value-type="string"/>
              </code>
        </time-aligned-layer>
     </coding-file
      ...
   </agent-codings>
</codings >
...
```

da-types.xml

```
...
<da-type nite:id="da_2" name="Questions">
    <da-type nite:id="da_21" name="q"
               gloss="Information Request"/>
    <da-type nite:id="da_22" name="qo"
               gloss="Open-ended question"/>
    <da-type nite:id="da_23" name="qh"
               gloss="Rhetorical question"/>
 </da-type>
...
```

m4-7.p2.word-segmentation.xml

```
...

 <word nite:id="m4-7.p2.word-segmentation.128" nite:start="266.466"
               nite:end="266.576" word="what"/>
 <word nite:id="m4-7.p2.word-segmentation.129"  nite:start="266.576"
               nite:end="266.776" word="kind"/>
 <word nite:id="m4-7.p2.word-segmentation.130"  nite:start="266.776"
               nite:end="266.946" word="of"/>
 <word nite:id="m4-7.p2.word-segmentation.131"   nite:start="267.026"
               nite:end="267.536" word="[AH]"/>
 <word nite:id="m4-7.p2.word-segmentation.132"   nite:start="267.536"
               nite:end="268.256"  word="preprocessing"/>
 <word nite:id="m4-7.p2.word-segmentation.133"  nite:start="268.256"
               nite:end="268.366" word="are"/>
 <word nite:id="m4-7.p2.word-segmentation.134"  nite:start="268.366"
               nite:end="268.476" word="you"/>
 <word nite:id="m4-7.p2.word-segmentation.135"  nite:start="268.476"
               nite:end="268.906" word="using."/>
...
```

m4-7.p2.dialog-act.xml

```
...

<dact nite:id="m4-7.p2.dialog-act.15" addressee="p0">
          <nite:pointer role="da-aspect"  xlink:href="da-types.xml#xpointer(id(da_21))" xlink:type="simple"/>
          <nite:child xlink:href="m4-7.p2.word-segmentation.xml#xpointer
           (id('m4-7.p2.word-segmentation.128')/range-to(id('m4-7.p2.word segmentation.135')))
xlink:type="simple"/>
 </dact>
...
```

Figure 3.1: An example of the M4 corpus

The dialogue act layer is defined as a structural layer in which annotations have children drawn from the word layer. The word layer, on the other hand, is defined as a time-aligned layer in which elements have timing information. These two layers contain annotations of the same type. However, layers may contain annotations drawn from a set of related types that together cover an observation. For example, a transcribed speech can be covered with words, vocal sounds (e.g. laughs and coughs) and so on. Annotations layers are further organized into codings ("dialogue-act" and "word-segmentation"). Although each of the presented codings consists of a single layer, codings in general may contain sequences of layers where each layer takes children from the next layer in the sequence ending either with a layer that does not have children or with a layer that is the top layer of another coding (Carletta et al., 2003). All layers in a coding are either for a particular agent (agent coding) or for the interaction as a whole (interaction coding). Both codings described here are agent codings. For agent codings, there is one coding file per agent. Figure 3.1 shows the dialogue-act and word-segmentation codings for agent "p2". Furthermore, codings for each observation - each meeting in our case - are stored in separate files. The example presented in Figure 3.1 is an excerpt from the meeting named "m4-7". It is to be noted that the dialogue act ontology as a corpus resource is stored in a separate XML file that can be referenced from codings ("da-types.xml"). The links between files are specified using XLink syntax.

This kind of representation has a number of advantages that are listed in (Carletta et al., 2003, 2005b). One of the advantages of using NXT storage format is that it enables data to be further processed and analyzed using a query language integrated in NXT - NXT Query Language (NQL)(Evert and Voormann, 2002; Carletta et al., 2005b). NXT provides a query library - NXT Search - that is used to evaluate queries expressed in NQL. The results are returned as list structures that can be further processed. Furthermore, query results can be saved in an XML format as well as in the Excel spreadsheet format. The query library also provides a graphical user interface, NXT Search GUI, that enables search over a corpus stored in the NXT format.

The AMI and M4 meeting corpora were created using a set of distinct annotation tools, some of which were developed using NXT[2] (Reidsma et al., 2005). Remaining tools store data in different formats. Annotations produced using these tools were translated in the NXT stand-off XML format.

3.2 Reliability

Krippendorff (1980) proposed three tests for assessing reliability of annotated data: stability (intra-annotator reliability), reproducibility (inter-annotator reliability) and accuracy.

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

[2] tools are freely available as a part of NXT

annotation. It is measured by giving several annotators the same data to annotate independently, following the same coding instructions. *Accuracy* is the degree to which an annotator's judgments conforms to a known standard. It is measured by comparing annotations of one coder with what is known to be standard. A standard can be, for example, the annotation produced by the schema's designer. Assessing the accuracy against such a standard, is useful for testing the performances of annotators in the course of training. However, in many cases the standards against which the accuracy should be established are not available. Therefore, for obtaining valid research results data on which they are based should be "*at least reproducible*, by independent researchers, and at different times, using the same instructions for coding the same set of data" (Krippendorff, 1980, p.132).

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 (Krippendorff, 2004b). Krippendorff's Alpha (α) has been suggested as a more suitable measure (Krippendorff, 1980).

Kappa and Alpha are chance-corrected agreement coefficients that measure the extent to which observed agreements among annotators differ from random agreement, that is, agreement by chance. Both agreement coefficients can be expressed in α's general form[3]:

$$\text{Agreement} = 1 - \frac{\text{Observed Disagreement}}{\text{Expected Disagreement}} = 1 - \frac{\mathrm{D}_o}{\mathrm{D}_e} \tag{3.1}$$

When $D_o = 0$, Agreement=1; when $D_o = D_e$ Agreement=0. As a result, D_e is the zero point of this agreement measure. Furthermore, if the observed disagreement is greater than the expected disagreement ($D_o > D_e$), Agreement$<$ 0. Kappa and Alpha differ in a way the expected disagreement is calculated. For a detailed description on how to calculate α and κ expected disagreements, we refer to (Krippendorff, 2004b).

To assess the reliability of most of the annotation schemas presented in this thesis, we applied both agreement coefficients. As the obtained Kappa and Alpha values were nearly identical, we report only Kappa values. The Alpha coefficient has been used for assessing the reliability of annotation schemas that take into account *partial* agreement between annotators.

For the evaluation of Alpha and Kappa values, the Krippendorff's (1980) scale has been adopted as standard in the discourse and dialogue processing community. 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. However, there are also some less strict scales used for inferring reliability from agreement measures in various fields (Landis and Koch, 1977; Fleiss, 1981). According to the cited scales an agreement coefficient below 0.40 denotes poor agreement. All these scales were proposed with considerable hesitations. From all these, it may be concluded that there is no single threshold against which agreement values can be measured

[3] κ's canonical form is $\kappa = \frac{P(A)-P(E)}{1-P(E)} = \frac{\text{Observed Agreement-Expected Agreement}}{\text{1-Expected Agreement}}$

in order to estimate reliability of any annotation schema. Krippendorff (2004a, p.242) claims that "the choice of reliability standards should always be related to the validity requirements imposed on the research results, specifically to the cost of drawing wrong conclusions". Currently, novel approaches to reliability analysis are being examined that focus more on the analysis of the use of the annotated data than to biasing reliability estimation to a particular agreement coefficient and inferring data reliability from it. In this chapter, we present reliability analysis of annotation schemas performed in the traditional way.

3.3 The M4 corpus

In this section, we present the M4 corpus that was primarily designed for studying addressing behavior in face-to-face meetings. First, we describe the corpus design where we present the meetings collection and annotation scheme. Then we present the analysis of the reproducibility and stability of the annotation scheme. A set of the corpus' annotations of the M4 meetings is available as a part of the M4 meeting collection[4].

3.3.1 Meeting data

The corpus consists of 12 meetings recorded at the IDIAP smart meeting room (Moore, 2002). The meeting recordings are available through the Media File Server[5]. The IDIAP meeting room is equipped with fully synchronized multi-channel audio and video recording devices (see Figure 3.2). Of the 12 meetings, 10 were recorded within the scope of the M4 project. In these meetings, participants were gathered together to discuss personal experiences regarding various issues such as a book, a movie or a trip. The meetings were scripted as to which actions the participants should undertake, but not as to what they should 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. In addition, one of the AMI pilot meetings recorded at the IDIAP meeting room is included in our corpus. 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.

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

3.3.2 Annotation schemes

The M4 scheme contains annotations of dialogue acts, adjacency pairs, addressees and gaze direction. We also considered annotation of deictic hand gestures but instances of

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

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

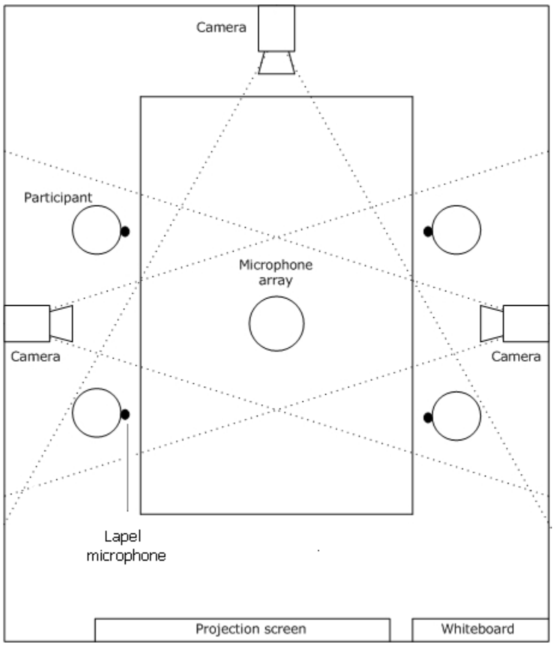


Figure 3.2: The configuration of the IDIAP meeting room (M4 data collection)

this type were found to occur very rarely in the data. Annotations of meeting actions that were made available to us, were converted into the NXT format and added to the corpus[6].

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 the M4 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.

The dialogue act tag set employed for the corpus creation, the M4 dialogue act tag set, is based on the MRDA (Meeting Recorder Dialogue Act) set Dhillon et al. (2004). The MRDA tag set represents a modification of the SWDB-DAMSL tag set (Jurafsky et al., 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 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.

Unlike MRDA, each utterance in the M4 annotation scheme is marked as Unlabeled or with exactly one tag from the tag set that represents the most specific utterance function.

[6] annotations of meeting actions stored in the ANVIL XML format are available http://www.idiap.ch/mmm/corpora/m4-corpus/m4_annotations

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.

The M4 dialogue act tag set was 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 that are objective and factual statements as well as utterances that are opinions and other subjective statements.
- **Acknowledgments and Backchannels**
  - **Acknowledgement** [MRDA: Acknowledgement, Backchannel]
    The Acknowledgment tag is a common tag used for acknowledgments and backchannels. Acknowledgments 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 acknowledgments 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 asked in a rather 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 that exhibit a (partial) disagreement to or a (partial) rejection 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 that 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 that influence the listeners' communicative or non-communicative future actions such as commands, suggestions, proposals.
    - **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 listeners (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 has 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 Unlabeled. 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.

Adjacency pairs

As discussed in Chapter 2, adjacency pairs are 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.

In the M4 schema, adjacency pairs are labelled at a separate level from dialogue acts. Labelling of adjacency pairs, as a form of relational labelling, consists of marking dialogue acts that occur as their A-part (marked as the *source*) and B-part (marked as the *target*). If a dialogue act is an A-part with several B-parts, for each of these B-parts, a new adjacency pair is created.

Addressees

Addressee annotation in the M4 schema is concerned with distinguishing whether the whole group, a subgroup of participants or an individual is addressed and in the case of subgroup or individual addressing which subgroup or individual it is. The schema specifies several sources of information that should be taken into account while determining who is being addressed: (1) content and type of dialogue act performed by the speaker, (2) conversational situation including, for example, information as to who has been previously talking, (3) the speaker gaze, postures and gestures and (4) meeting context relating to types of group activities participants perform in the meeting. Furthermore, when the speaker is indirectly addressing an individual while talking to another individual or to a group, the individual the speaker is talking to or the group is marked as the addressee. This relates to distinction between the target and addressee roles presented in Section 2.6.5.

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

- a single participant: $\mathbf{P}_x$
- a subgroup of participants: $\mathbf{P}_x$,$\mathbf{P}_y$
- the whole group: $\mathbf{P}_x$,$\mathbf{P}_y$,$\mathbf{P}_z$
- **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 Unlabeled.

Gaze direction

Annotation of gaze direction involves two types of activities: labeling the changes in the gazed targets and labeling the gazed targets themselves. 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 annotation schema does not impose any requirements regarding the precision of marking the changes in the gazed target.

Meeting actions

For labelling group actions in meetings the following set of meeting actions has been employed[7]:

- **monologue** - one participant speaks continuously without interruption
- **discussion** - all participants engage in conversations
- **presentation** - one participant at the front of the room makes a presentation using the projector screen
- **whiteboard** - one participant at the front of the room talks and makes notes on the white-board

[7] see http://www.m4project.org/publicDelivs/D1-2.pdf

- **consensus** - all participants express consensus
- **disagreement** - all participants express disagreement
- **note-taking** - all participants write notes

The listed definitions of meeting actions are taken from (McCowan et al., 2003).

3.3.3 Assessment of reliability of the M4 annotation schemas - an overview

To measure reliability of the annotation schemas, we have performed stability and reproducibility tests (see Section 3.2). To estimate reliability of dialogue act, addressee and gaze annotation we applied the Kappa test. 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.

3.3.4 Inter-annotator reliability

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 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.

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.

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 adjacency pairs.
- **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 Unlabeled 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 3.1 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 3.1: Segmentation agreement (R-C pair: Reference annotator (R)-Comparison annotator (C))

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.

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 3.2. According to Krippendorff's scale annotators in each DA group reached an acceptable level of agreement.

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 3.2: 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 (Krippendorff, 1980). Single-category reliability was measured by grouping the remaining categories into one category and measuring the agreement among annotators regarding the assignment of units to these two categories. Table 3.3 shows the results of performing single-category tests for only one R-C pair in each DA group.

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 chance: in all cases where one annotator identifies one of these two categories, the other annotator does not. The results show a very 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 Acknowledgment. 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.

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

Category	**B-E**	**M-R**
Statement	0.82	0.72
Acknowledgment	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 questions	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

Table 3.3: Single-category reliability for DA tags (Kappa values)

Kappa values for addressee annotation are shown in Table 3.4.

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 3.4: Inter-annotator agreement on addressee annotation

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 labeled 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.

Reliability of adjacency pairs annotation

According to the M4 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 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 (L) that is compounded of its A-class and B-class ($L = A - class | B - class$).

Given a list of dialogue acts $DA = [da_1, \ldots, da_n]$, 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 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 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 ($\langle 0, \{\} \rangle$).

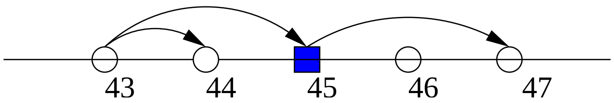


Figure 3.3: A graphical representation of the context of dialogue act 45. The label that encodes this context is $\langle 0, \{2\} \rangle \,|\, \langle -2, \{1, 2\} \rangle$

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 3.5. Figure 3.3 illustrates the relation between the context of the dialogue act 45 and the AP label that encodes this context.

DA	$\mathbf{C}_1$	$\mathbf{C}_2$	$\mathbf{C}_1(1)$	$\mathbf{C}_2(1)$
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 3.5: An example of adjacency pairs annotation (C_1 and C_2: original AP annotations; $C_1(1)$ and $C_2(1)$: AP labels)

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: $\delta = 0$ 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 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 [8]:

$$sim(c_1, c_2) = \frac{2|c_1 \cap c_2|}{|c_1| + |c_2|} \tag{3.2}$$

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

$$\delta'(\langle -n_1, O_1\rangle, \langle -n_2, O_2\rangle) = 1,\ n_1 \neq n_2 \tag{3.3}$$

$$\delta'(\langle -n, O_1\rangle, \langle -n, O_2\rangle) = 1 - sim(O_1, O_2) \tag{3.4}$$

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

$$\delta_\lambda(L_1, L_2) = \lambda \cdot \delta'(A_1, A_2) + (1-\lambda)\delta'(B_1, B_2), \tag{3.5}$$

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

[8] The defined similarity measure is known as *Dice coefficient* Manning and Schutze (1999)

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 marks 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.

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.

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.

3.3.5 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 3.6 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.

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 3.6: Intra-annotator agreement

3.4 The AMI corpus

The AMI meeting corpus is a multi-modal corpus consisting of 100 hours of meeting recordings. For a detail description of the corpus, we refer to (Carletta et al., 2005a). A large proportion of the AMI data collection contains scenario-driven meetings involving design teams focused on the development of a remote control prototype. The corpus also contains natural meetings collected mostly for validation and generalization of findings from the elicited data. To facilitate research in a wide range of research communities, the AMI meetings were annotated for many different phenomena including speech transcription, dialogue acts, addressee, topics, named entities, summaries, argumentation, emotions, individual actions and focus of attention.

As previously discussed, only a small part of the AMI scenario-based data was annotated with addressee information. In this section, we, therefore, describe the AMI scenario-based collection. Then, we give an overview of those annotation schemas that had been used for annotation of phenomena employed as features in our addressee classification models. Finally, we report results on reliability analysis of those annotation schemas.

3.4.1 Meeting data - scenario-based collection

The AMI meetings were recorded in three meeting rooms constructed at different project sites: IDIAP, the University of Edinburgh, the TNO institute. The rooms are relatively similar but differ in shape, construction and configuration of recording equipment (see Carletta et al., 2005a). The data were captured using a wide range of fully synchronized multi-modal recording devices including close-up and room view cameras, close-talking and far-field microphones, and instruments that capture data presented in meetings such as slides, the whiteboard content and individual note-taking. The rooms were also equipped with artifacts such as laptops or remote control prototypes.

According to the AMI meeting scenario (Post et al., 2004), meeting participants, who play roles of employees in an electronic company, constitute a design team responsible for the development of a new remote control prototype. They are assigned to roles of project manager (PM), industrial designer (ID), user interface designer (UI) and marketing expert (ME) with different types and degrees of responsibility regarding the design project as a whole. A design process is carried out through a series of four meetings each of which representing one phase in the design process: project kick-off, functional design, conceptual design and detailed design. Meetings are on average approximately 30 minutes long; each meeting consists of 4 participants.

3.4.2 Annotation schemas

This section overviews several AMI annotations schemas that are employed for annotating phenomena relevant to addressing: dialogue acts, addressee, named entities, topics and focus of attention. All these schemas are available at the AMI Meeting Corpus public release website[9]. The AMI individual action schema specifies annotation of hand and head deictic gestures. Empirical analysis of the data have shown that these types of gestures occur very rarely in the data. Moreover, annotation of hand gestures for most of the meetings annotated with addressee information are not yet available.

Dialogue acts and addressees

Segmentation Transcribed speech is segmented into dialogue act segments according to the speaker's intention. If within a turn the speaker performs, for example, two subsequent questions each question denotes different intentions and therefore is marked as a separate segment. Similarly, if during a speech, a change in addressee occurs the speech is segmented into two segments each of them denoting different intentions. Although collaborative dialogue acts - dialogue acts produced by two or more speakers - represent a single intention, they are split over segments assigned to each speaker. Each of these segments is marked with the same dialogue act and addressee label.

Dialogue act tag set The AMI dialogue act tag set is a one-dimensional tag set: each dialogue act segment is marked with exactly one tag from the set. The tags in the AMI schema are grouped into several classes as follows:

- **Special classes to allow complete segmentation** - the AMI schema imposes that the transcription is completely segmented. However, not everything present in the transcript conveys speakers' intention, that is, not everything is considered as a dialogue act. The segments that do not convey speaker intention are labelled with one of the following tags:
  - **Backchannel** - marks utterances produced in the background by someone who has just been listening to a speaker, expressing merely that the listener is following the conversation.
  - **Stall** - this tag pertains to mechanisms of grabbing or maintaining the floor. Stalls are also used to mark valid starts to what the speaker actually wants to say ("That is — that is really unacceptable").
  - **Fragment** - marks those segments that do not convey a speaker intention but that are neither Stalls nor Backchannels. They include, among other things,

[9] http://www.idiap.ch/amicorpus/documentations/guidelines-1/

incomplete contributions that do not convey a speaker's intentions (e.g "I think we").

- **Acts about information exchange**
  - **Inform** - the Inform act is employed by a speaker to give information.
  - **Elicit-Inform** - marks the speaker's request for some other participant(s) to provide some information.
- **Acts about possible actions**- this group of dialogue acts refers to possible actions that could be performed by a meeting participant or a group of participants, or some person or group in the wider environment (e.g marketing team of the company).
  - **Suggest** - marks the speaker's intention regarding the actions of another individual or to the actions a group in the meeting or in the wider environment.
  - **Offer** - marks the speaker's intention regarding his own actions.
  - **Elicit-Offer-Or-Suggestion** - marks the speaker's request for someone present at the meeting or for some person or a group in the wider environment to make an offer or suggestion.
- **Commenting on previous discussion**
  - **Assess** - marks any evaluation of something the group is discussing. The evaluation can refer to another dialogue act, to something present in the meeting room or to some actions performed by members of the group. Assessments include, among other things, agreement or disagreement with something previously said, acceptance or rejection of an offer, or any opinion about information provided.
  - **Comment-About-Understanding** - is used for those kinds of comments in which the speaker indicates that he did or did not understand or hear what a previous speaker said.
  - **Elicit-Assessment** - marks the speaker attempts to elicit assessments about something previously discussed or done.
  - **Elicit-Comment-About-Understanding** - marks the speaker attempts to elicit a comment about whether something that was previously said or done has been understood.
- **Social acts** - this group contains acts that do not contribute to the overall discussion but rather have a social impact, that is, have an affect on the interpersonal relationship in the group.
  - **Be-Positive** - marks those social acts that express positive feelings towards an individual or the group and in that way improve relationships in the group itself.

  - **Be-Negative** - marks those social acts that express negative feelings towards an individual or the group (e.g. hostile jokes, criticisms).

- **Bucket class**
  - **Other** - is used to mark all other dialogue acts that cover the speaker intention but do not fit any other categories in the AMI tag set. They include, among other things, self-addressed speech.

Relations The AMI schema imposes labelling of the responsive aspects of dialogue acts. Dialogue acts can be a response to something someone previously has said or done or to something present in the environment. The responsive aspect of a dialogue act is labelled by marking a relationship with the dialogue act as its target. The source of the relationship can be a dialogue act performed by another speaker (e.g. Elicit-Inform - Inform) or by the same speaker (e.g. when the speaker is assessing his own contribution). When a dialogue act is related to something that is not expressed verbally or to more than one previous dialogue act, the source of the relationship is not specified.

In addition to marking the source and the target of a relationship, the schema specifies labelling of a type of the relationship as Positive, Negative Partial or Uncertain based on whether the target supports, rejects, partially supports or expresses uncertainty about the source, respectively.

Reflexivity Dialogue acts in the AMI schema are also distinguished as to whether they are 'reflexive' or not. A dialogue act is marked as reflexive if its content is about how the group carries out the task. Examples include acts concerned with how to conduct the meeting (e.g. a suggestion to start a meeting with a presentation) or how to divide the work (e.g. a suggestion that the user interface designer and the industrial designer should cooperate).

Addressees The addressee in the AMI schema is labelled for each dialogue act that is not mark as Backchannel, Stall, Fragment or Other. Addressee annotation is concerned with distinguishing whether an individual or a group is addressed and in the case when an individual is addressed which individual it is. The schema does not make a distinction between addressing a subgroup of participants or addressing the whole group. In both cases, dialogue acts are marked as addressed to the group. Similarly to the M4 schema, when a speaker is indirectly addressing an individual while talking to another individual or to a group, the individual to whom the speaker is talking or the group is marked as the addressee.

The addressee tag set consists of the following tags:

- **Individual**: **A**, **B**, **C**, **D**

- **Group** - is used when more than one participant is addressed

- **Unclassifiable** - is used when annotators cannot determine who is being addressed

Participant's labels A, B, C and D are related to speech channels in the AMI corpus.

Named Entity

The AMI named entity annotation schema is based on NIST "1999 Named Entity Recognition Task Definition"[10] with some modifications made for the particular needs of the AMI corpus. It codes entities (people, locations, organizations, artifacts), quantities (money, measures, percents, numbers), temporal information (dates, times, durations), materials, shapes and colours. The named entity tags are organized into the following classes: **Enamex**, **Timex**, **Numex**, **Artifact**, **Colour**, **Shape**, **Materials**.

Out off all categories, Enamex is the only category employed for the experiments on addressee identification present in this thesis. It comprises several subcategories that mark references to named persons or family (**Person**), name of politically or geographically defined locations (**Location**) and organizational entities (**Organization**). In the AMI schema, the **Person** category is further subdivided into the **Participant**, **Experimenter** and **Other_Person** categories. The Participant category consists of four tags that refer to participant roles in the AMI scenario: **Project_Manager**, **Interface_Specialist**, **Marketing** and **Industrial_Designer**. The Experimenter category marks references to anyone in the research team who gives instructions to the participants or sets up recording devices. Any other person referred to by participants is marked with the Other_Person category. Regarding the Person category, not only proper names, but also other types of designators to individuals are annotated. For example, in the scenario meetings all occurrences of scenario roles are marked with the corresponding Participant category.

Topic annotation for scenario-based meetings

The AMI topic annotation schema specifies segmentation of meetings into coherent topic segments as well as labeling of those segments with a proper topic description.

Topic segments in the AMI schema can be nested: a topic may contain subtopics that are related to that topic. Each topic segment is labeled with a short topic description. Although the schema offers a predefined set of topic descriptions that are expected to occur in the scenario-based meetings, annotators were also allowed to introduce a new topic description, if needed. The standard topic descriptions are divided into three categories:

- **Top-Level Topics** refers to topics whose content mainly reflects the meeting structures defined according to the scenario. Examples include "project specification and roles of participants", "marketing expert presentation", "drawing animals", "discussion" or "evaluation of project process".

- **Sub-Topics** refer to more specific topics that top-level topics are made of. They reflect largely things the group is discussing such as "look and usability", "existing products", "user requirements" or "project budget".

[10] http://www.nist.gov/speech/tests/ie-er/er_99/doc/ne99_taskdef_v1_4.pdf

- **Functional** topics refer to those parts of meetings that are not real topics as they do not contain useful information regarding the design process itself. They reflect either the actual process and flow of meetings ("opening", "closing", "agenda/equipment issues") or they are irrelevant ("chitchat"). Functional topics can be either top-level topics or subtopics, but they are not further refined into subtopics.

Focus of Attention (FOA)

FOA annotation reflects annotation of gaze direction of meeting participants. It consists of labeling of changes in the gazed target as well as labeling of gazed targets themselves. The gazed targets are labeled with the following tags:

- one tag for each participant: **PM**, **ID**, **UI**, **ME** [11]
- **Whiteboard**
- **Slide-screen**
- **Table** - is used when participants are looking at something on the table such as their hands, documents or the laptop
- **Unspecified** - is used when the participant is not focusing on anything in particular or focusing on something that is not covered with the other tags (e.g. the bookshelf).

In the case of ambiguity between focusing on a participant and focusing on the slide screen or the whiteboard, the participant is labeled as focus of attention. Changes in the gazed targets are labeled with a high precision; changes are marked in the middle of eye movement between old and new target.

3.4.3 Reliability of annotation schemas

Reliability analysis of the AMI schemas is ongoing work carried out by various project sites. For most annotation schemas, the results have not yet become available. Up to now, only the reliability analysis of the named entity annotation schema has been published. In this section, we present analysis of reliability of the addressee and FOA annotation[12].

Reliability of addressee annotation

The reliability of the addressee annotation schema was evaluated by comparing the sets of annotations of one of the AMI meetings provided by four annotators. Inter-annotator agreement was measured for each pair of annotators using the Kappa coefficient. Although one of the trained annotators (annotator *d* in Table 3.7) did not take part in the corpus

[11] only scenario based meetings were annotated with focus of attention

[12] This analysis is based on joint unpublished work with Dennis Reidsma and Rieks op den Akker

creation, we made use of annotations provided by this annotator to evaluate the reliability of the annotation schema.

As discussed in Section 3.4.2, addressees in the AMI schema are labeled for each dialogue act segment that is not marked as Backchannel, Stall, Fragment or Other. Agreement on addressee classification was thus measured over those dialogue act segments for which both annotators agreed on their boundaries and neither of the annotators labeled them as Backchannel, Stall, Fragment or Other. Furthermore, we performed a reliability test for the real addressee tags which mark individual (A, B, C, D) or group addressing. Hence, segments marked with the Unclassifiable addressee tag by any of two annotators were also excluded from the analysis. The results of the test are summarized in Table 3.7.

Annotators pair	**N**	**ADD(κ)**
a-b	353	0.63
a-c	301	0.48
a-d	427	**0.67**
b-c	349	0.50
b-d	400	0.59
c-d	340	**0.45**

Table 3.7: Inter-annotator agreement on addressee annotation. N -number of agreed segments. Total number of segments per annotator: a:814, b:715, c:701, d:816. Addressee tag set: A, B, C, D, Group

The results indicate an overall low agreement between annotators on addressee annotation on the AMI data. Annotator *c*, as the result shows, is the most problematic annotator. Inspection of the confusion matrices for this annotator has shown that he tends to use the Group category more than the other annotators. In general, all annotators had difficulties in determining whether a dialogue act was addressed to a group or to some individual. When annotators agreed that an individual was addressed in almost all cases they agreed who that individual was. We will discuss the impact of this type of confusion on the performances of addressee classifiers in Chapter 5. It is to be noted that the same type of confusion has been observed for addressee annotation on the M4 data (see Section 3.3.4). The addressee annotators had most difficulties distinguishing addressees of those dialogue acts that were marked as Inform or Assess.

We have also measured agreement between annotators on addressee annotation using the complete addressee tag set that includes the Unclassifiable category. As presented in Section 3.4.2, this category is used when annotators cannot decide who is being addressed by the speaker. Informal inspection of the confusion matrices have shown that annotators hardly agree on labelling a dialogue act with this category which leads to a significant decrease of Kappa values for each pair of annotators ($0.33 \leq \kappa \leq 0.57$). Since most of the confusions were between individual and group addressing, we can conclude that the Unclassifiable category was used when annotators were unsure whether a group or an individual was addressed. Due to high disagreement between annotators on this category, we may also conclude that the use of the Unclassifiable category reflects more the annotators' personal opinion about his uncertainty than some class of items that was difficult to annotate with addressee information.

Reliability of FOA annotation

Inter-annotator agreement on FOA annotation was estimated among six annotators on the annotations produced for one meeting participant (PM) in one of the AMI meetings. The approach applied for measuring agreement first aligns segments for a certain threshold to adjust differences in annotators' perceptions of the timing of changes in the gazed targets. Then, it estimates annotators agreement on FOA labeling over the aligned segments. Informal visual inspection of the results of the alignment has shown that aligned segments are indeed sensible in a sense that they represent the same event. Various thresholds were considered in the alignment procedure ranging from 0.1s to 0.8s. It was observed that for thresholds higher than 0.5s the number of aligned segments do not change significantly; for some annotator pairs, the segments were already pretty well aligned at the threshold 0.3s.

Table 3.8 summarizes the percentage agreement on segmentation for each pair of annotators for the alignment threshold of 0.5s.

	a	**b**	**c**	**d**	**e**	**f**
a	–	91%	91%	85%	80%	95%
b	77%	–	84%	79%	64%	87%
c	71%	77%	–	75%	61%	79%
d	72%	80%	82%	–	68%	77%
e	77%	74%	76%	77%	–	76%
f	75%	82%	80%	72%	62%	–

Table 3.8: Percentage agreement on segmentation - (x,y) entry denotes the percentage of segments for annotator x that were aligned with segments of annotator y

Agreement on FOA labelling was estimated on the aligned segments using the Kappa measure. The pairwise Kappa values range between 0.84 and 0.95 which indicates very good agreement on FOA labeling.

Reliability of named entity annotation

Reliability of the AMI named entity annotation schema is reported in Karaiskos (2005). Segmentation agreement was measured over all text segments which any of the annotators marked as a named entity. The reported result showed that in 90.1% of cases annotators fully agreed on segment boundaries whereas in 91.3% of cases annotators agreed on at least one word of a segment. Furthermore, annotators reached very good agreement on named entity classification ($\kappa = 0.98$, N = 374) that was measured over those named entity segments for which annotators agreed on at least one word.

3.5 Comparison between the M4 and AMI meeting corpora

In this section, we compare the M4 and AMI meeting corpora with respect to meeting data collections as well as with respect to annotation of phenomena that were annotated on both data sets. Regarding the AMI corpus, we focus only on the scenario-driven part of the data collection.

Data collection

The M4 corpus consists mostly of the short - approximately 5 minutes long - discussion meetings recorded in the context of the M4 project whereas the AMI corpus contains longer - approximately 30 minutes long - task-oriented meetings, where the task is the design of a new remote control prototype. Each meeting is concerned with a particular phase in the design process that is carried out through the series of four meetings. The M4 meetings are scripted at the level of the sequence of group activities that participants perform, but content is natural and unconstrained. The AMI meetings, on the other hand, are more natural meetings structured according to the AMI scenario.

The M4 and AMI meetings consist of 4 participants. In contrast to the the M4 meetings, participants in the AMI meetings are assigned particular roles. Meetings from the M4 collections were recorded in the IDIAP room whereas meetings from the AMI collection were recorded in the IDIAP, TNO and Edinburgh instrumented meeting rooms. Although the layouts of the IDIAP room in the M4 and AMI collection are similar, the room in the AMI collection was equipped with additional recording equipment and attention distracters (e.g. laptops, a remote control prototype). In all three rooms, the seating arrangement includes two participants at each of two opposite sides of the table.

Dialogue act annotation

Table 3.9 summarizes the mapping between the dialogue tag sets employed for the creation of the M4 and AMI meeting corpora.

The Fragment tag in the AMI schema is used when the speaker starts saying something but did not get far enough to express his intention because he was either interrupted or he changed his mind and said something else or stopped altogether. Interrupted or abounded incomplete contributions of this kind are marked as Unlabeled in the M4 schema. However, if the speaker continues talking, the incomplete contribution that is considered as the “false start” is merged with the segment that follows. Similarly, utterances marked as Stall in the AMI schema are marked as Unlabeled in the M4 schema if they occur alone, that is, if the speaker does not take the floor and continue talking. In all other case, they are not segmented into separate segments.

M4 DA Tag Set	AMI DA Tag Set
Statement	Inform, Assess, Be-Negative, Be-Positive
Information-Request	Elicit-Inform, Elicit-Assessment
Open-ended Question	Elicit-Offer-Suggestion, Elicit-Assesment
Rhetorical Question	Other
Acknowledgement	Backchannel, Other
Assessment/Appreciation	Assess, Be-Positive, Be-Negative
Positive Response	Inform, Assess
Negative Response	Inform, Assess
Uncertain Response	Assess
Influencing-listeners-action	Suggest
Committing-speaker-action	Offer
Follow Me	Elicit-Comment-About-Understanding
Repetition Request	Comment-About-Understanding
Understanding Check	Elicit-Inform, Elicit-Assessment, Comment-About-Understanding
Apology	Be-Positive
Thanks	Be-Positive
Other polite	Be-Positive
Unlabelled	Fragment, Stall, Other

Table 3.9: Mapping between the M4 and AMI dialogue act tag sets

Addressee annotation

As the AMI addressee annotation schema does not make a distinction between addressing a subgroup of participants and addressing the group as a whole, the Group addressee tag is employed for marking that a group is addressed regardless of the number of participants in the group. The M4 schema, on the other hand, contains separate tags for addressing a subgroup and addressing the whole group.

Labeling relations between dialogue acts

Labeling relations between dialogue acts in the M4 schema is considered with labeling adjacency pairs as classically defined. As adjacency pairs represent units that contain paired utterances produced by different speakers, a dialogue act of one speaker is marked as related to a dialogue act of another speaker or as not being related at all.

The AMI schema, on the other hand, encompasses a more general notion of relation between dialogue acts that reflects different kinds of responding aspect of a dialogue act, namely: a dialogue act of one speaker can be related to a dialogue act of another speaker (e.g. adjacency pairs), or to a dialogue act of the same speaker (e.g. rhetorical relations) or even to something that is not expressed verbally. In the later case, the source of the relation is left unspecified. In addition to marking the source and the target of a relation, the AMI schema requires a type of relation to be classified as positive, negative, partial or uncertain.

Gaze annotation

Gaze direction of participants in the M4 meetings was labelled using the simple set that includes one tag for each individual and NoTarget. In the AMI schema, the NoTarget category is further refined in the table, whiteboard, slide-screen and unspecified categories. Furthermore, in the AMI schema, person A is considered as looking at B, only when A looks at the face of B. The M4 schema, on the other hand, does not impose this kind of restriction when labeling an individual as a gazed target: looking at a participant's hands or back is marked as looking at that participant. In contrast to the M4 schema, the AMI schema imposes the requirement for a high precision on labelling changes in the gazed targets.

3.6 Summary

In this chapter, we presented the M4 and AMI corpora that were employed for the experiments on addressee identification described in Chapter 5. Regarding the AMI corpus, we focused only on those parts that have been used in our study. Three aspects of the corpora were discussed and compared: data collection, annotation schemas and reliability of the annotation schemas.

Reliability analysis of the annotation schemas have shown that annotators involved in the design of both meeting corpora were able to reproduce the gaze labeling reliably. The annotations of dialogue acts and addressees on the M4 data were somewhat less reliable but still acceptable. The annotators involved in the design of the AMI corpora achieved lower agreement on addressee annotation than the annotators involved in the design of the M4 corpora. An important overall conclusion regarding addressee annotation is that human observers of meeting conversations mostly had difficulties distinguishing individual from group addressing. Analysis of subgroup addressing on the M4 data have shown that instances of this type occur rarely in the data and annotators failed to agree on subgroup categories. On the AMI data, annotation of subgroup addressing was not considered due to the same reasons. Therefore, subgroup addressing was not taken into account in developing models for automatic addressee identification.

Chapter 4

Bayesian Networks as Classifiers

We consider the problem of identifying the addressee of the speaker's dialogue act as a classification problem where classes are possible addressee values. There are numerous techniques for classification such as decision trees, neural networks or support vector machines. To model our problem domain, we have chosen Bayesian Networks (BNs). BNs are a class of models that graphically encode probabilistic relationships among a set of variables. They are a well-known tool for knowledge representation and reasoning under uncertainty. As such, BNs have been successfully exploited in numerous expert systems and decision support applications. In the context of classification, they have been applied for various tasks similar to addressee identification such as speaker detection or dialogue act recognition (Keizer and op den Akker, 2006; Rehg et al., 1999; Choudbury et al., 2002). In these tasks, multiple sources of information are combined. One advantage of BNs in comparison to other classification models is that, when learned from data, they provide an understanding of the domain being modeled. In other words, they extract explicit knowledge from data reflected in probabilistic causal relationships between variables that represent the domain. Another advantage of BNs is that they allow for the combination of data and domain knowledge.

This chapter is organized as follows. Section 4.1 describes the problem of classification. Section 4.2 first provides a brief introduction to BNs and then overviews various techniques for learning and inference in BNs (Section 4.2). Section 4.3 focuses on the use of BNs as classifiers. It provides an overview of several Bayesian Network (BN) classifiers that were employed for the experiments on addressee classification presented in Chapter 5. Dynamic Bayesian Networks (DBNs) are presented in Section 4.4. The chapter concludes with a discussion on the methodology for the evaluation of classifiers' performances (Section 4.5).

4.1 Classification

Classification can be viewed as a process of assigning a predefined set of class labels $\mathcal{C}$ to data instances described in terms of a set of attributes or features $\mathcal{A}$. A model or a function $\mathcal{F} : \mathcal{A} \Rightarrow \mathcal{C}$ which maps a data instance into a corresponding class label is referred to as a *classifier*. If the function is unknown, a classifier can be induced from a data set $\mathcal{D}$ consisting of pre-classified instances $\mathcal{D} = \{(\boldsymbol{a_i}, c_i)|i = 1, \ldots, n\}$ where $\boldsymbol{a_i} \in \mathcal{A}$ represents a vector of attributes which describes the *i-th* instance and $c_i \in \mathcal{C}$ is its corresponding class label. In the literature, this set is usually referred to as a training set. The inferred classifier is then used for assigning the class label to new, so far unseen, data instances.

A key concept in classification is that of uncertainty. Classification deals with uncertainty in several ways. First, data instances may be derived from a process which is not completely known. Second, the instances of the real world are represented with a set of the selected attributes that represents an approximation of the real instances which leads to the loss of information in perceiving that problem. Moreover, there may be some unwanted anomalies, so-called noise, caused by imprecision in collecting the attributes (e.g. errors in recoding, measuring or labelling). Although the underlying process can be deterministic, there can be some pieces of information that are not accessible - so-called unobservable attributes - which have also been used by the deterministic function in determining the class labels. Due to uncertainty, the underling process is modeled as a random process whose outcome is a random variable $\mathcal{C}$. The random variable does not describe the actual outcome of the process but describes the quantified uncertainty of each possible outcome. A common framework used for the quantification and manipulation of uncertainty is probability theory.

Classification is concerned with making optimal decisions under uncertainty: the goal is to decide which of the class labels C_i is to be assigned to a particular data instance described with a set of attributes $\boldsymbol{a}$ based on the estimated probabilities of each possible outcome $P(C_i|\mathbf{a})$. Optimality is concerned with minimizing the chance of assigning $\boldsymbol{a}$ to a wrong class. If the goal is to minimize misclassification error, the class label with the highest conditional probability $P(C_i|\mathbf{a})$ is chosen (Duda and Hart, 1973). However, in many applications, the consequences of making wrong decisions may have different impacts regarding different class labels. Those impacts are formalized in terms of *loss functions* which is a measure of the loss λ_{kj} incurred by assigning the class label C_j instead of the correct class label C_k to the new instance $\boldsymbol{a}$. The optimal decision in this case is concerned with minimizing the expected loss, that is, with choosing C_j for which the expected loss defined as $\sum_k \lambda_{kj} P(C_k|\mathbf{a})$ is minimal. In this thesis, classification is concerned with minimizing misclassification error.

From all these, it follows that we need to have a model for determining the conditional class probabilities $P(C_i|\mathbf{a})$ in order to make optimal decisions. The classification problem can thus be divided into two stages: the *learning* stage in which the training data is used to learn a model for estimating $P(C_i|\mathbf{a})$ and the subsequent *decision* stage in which these probabilities are used to make optimal class assignments (Bishop, 2006).

BNs are probabilistic graphical models that represent a full joint probability distribution over a set of random variables. Therefore, they can be used to model the joint distribution $P(C, \mathbf{a})$ from which $P(C|\mathbf{a})$ can be inferred. As they can also be learned from data using various statistical techniques, BNs provide a framework for performing classification. In the following section, we introduce BNs and discuss various methods for learning and inference in BNs.

4.2 Bayesian Networks

4.2.1. Definition. A Bayesian Network[1] is a probabilistic model over a set of random variables $X = \{X_1, \dots X_n\}$ which consists of two parts:

> **Qualitative part** (G) - a directed acyclic graph (DAG) represented as an order pair G=(X,E) where
>
> - X is a set of random variables $\{X_1, \dots, X_n\}$ that form the nodes of the network.
> - E is a set of direct edges which encode influential links between the pairs of nodes they are connecting. An edge from X_i to X_j denotes that the node X_i directly influences the node X_j. X_i is called the parent of X_j. The set of all parents of the node X_j is marked as $Pa(X_j)$.
>
> **Quantitative part** (θ)- a set of conditional probability distributions assigned to each node in the network $\{P(X_i|Pa(X_i))|X_i \in X\}$. A conditional probability distribution assigned to the node X_i quantifies the effects that its parents $Pa(X_i)$ have on the node.

Each variable in a BN can have either a finite set of possible values (discrete variable) or an infinite number of possible values (continuous variable). In this chapter, we focus only on BNs over a set of discrete random variables. An example of BNs is given in Figure 4.1.

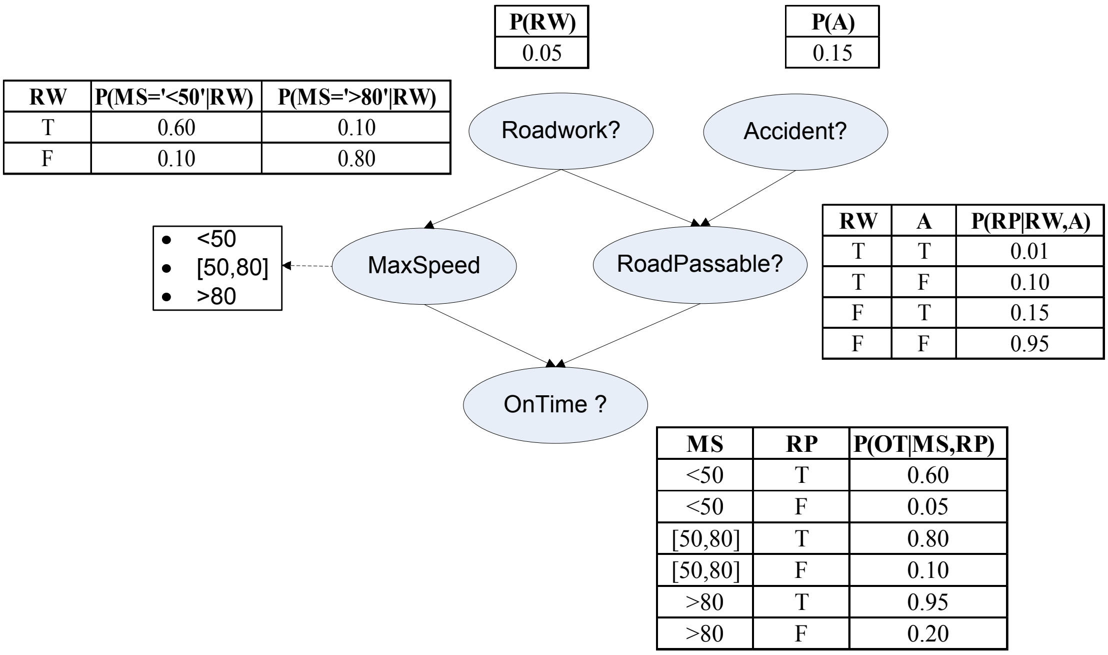
Definition 4.2.1 specifies the syntax of BNs. Regarding the semantic properties, BNs can be seen in two different ways. First, a BN can be seen as an encoding of conditional independence statements among the random variables represented with its nodes (*topological semantics*). Second, a BN can be viewed as a factorized representation of the full joint probability distributions over its random variables (*numerical semantics*).

Topological semantics

In the probability theory, two random variables X and Y are considered as conditionaly independent given the variable Z if $P(X|Y,Z) = P(X|Z)$. A relation between this quantitative notion of conditional independence and the qualitative properties of a BN is established in the concept of *d-separation*[2] (Pearl, 1988). Informally, the concept of d-separation is based on "blocking" the information flow between two variables X and Y in a network by some third variable Z. If the information flow between X and Y is blocked then changes in the certainty of X have no influence on changes in the certainty of Y, that is, X and Y are conditionaly independent given Z. One example of d-separation in the BN presented in Figure 4.1 is the following: if the values of *RoadPassable?* and *MaxSpeed* are known then information flow between *OnTime?* and *RoadWork?* is blocked; thus, *OnTime?* and *RoadWork?* are conditionally independent given *RoadPassable?* and

[1] also known as a **belief network** or a **causal network**

[2] '**d**'stands for '**directed**'



P(RW)
0.05

P(A)
0.15

RW	P(MS='<50'|RW)	P(MS='>80'|RW)
T	0.60	0.10
F	0.10	0.80

RW	A	P(RP|RW,A)
T	T	0.01
T	F	0.10
F	T	0.15
F	F	0.95

MS	RP	P(OT|MS,RP)
<50	T	0.60
<50	F	0.05
[50,80]	T	0.80
[50,80]	F	0.10
>80	T	0.95
>80	F	0.20

Figure 4.1: An example of Bayesian Networks

MaxSpeed. The same is valid for *OnTime?* and *Accident?*. The formal specification of d-separation can be found in (Pearl, 1988; Bishop, 2006).

There is also a more intuitive and simpler way to specify the topological semantics in BNs that is based on the notion of the *Markov blanket* (Pearl, 1988): a node in a BN is conditionally independent of all other nodes in the network given its Markov blanket which consists of the node's parents, children and children's parents. In other words, the Markov blanket of a node X is a set of nodes that isolates X from the rest of the graph. In our example of BNs, *Accident?* is independent of *MaxSpeed* and *OnTime?* given *RoadPassable?* and *Roadwork?*.

Numerical semantics

As previously mentioned, a BN over a set of random variables $\{X_1, \ldots, X_n\}$ represents a specific decomposition of a full joint probability distribution over the variables. The full joint probability distribution is given by the product, over all nodes in the graph, of a conditional probability distribution of each node given its parents in the graph:

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | Pa(X_i)) \tag{4.1}$$

A detailed explanation of the equivalence between topological and numerical semantics of BNs expressed through the d-separation criterion can be found in Bishop (2006).

The general formula for calculation of the full joint probability distribution over a set of random variables $X = \{X_1, \ldots, X_n\}$ is given by:

$$P(X_1,\ldots,X_n) = P(X_1)P(X_2|X_1)\ldots P(X_n|X_1,\ldots X_{n-1}) = \prod\nolimits_{i=1}^{n} P(X_i|X_1,\ldots,X_{i-1}) \tag{4.2}$$

It requires a large number of conditional probabilities to be estimated: assuming that all variables have at most k values, we need to assess at most $\mathcal{O}(k^n)$ numbers. The equation 4.1 provides more compact representation of the full joint probability distribution as it requires $O(nk^m)$ conditional probabilities to be estimated, assuming that each variable has at most $m \leq n-1$ parents.

Comparing the equation (4.1) with the general formula for the calculation of joint probability distribution (4.2), we see that the specification of the probability distribution provided by a BN is equivalent to general assertion if for every variable X_i the following holds.

$$P(X_i|X_1\ldots X_{i-1}) = P(X_i|Pa(X_i)) \tag{4.3}$$

given that $Pa(X_i) \subset \{X_1 \ldots X_{i-1}\}$ which can be obtained by labeling nodes in an ordering consistent with the order implicit in the graph structure. From this equation, it follows that a BN is a correct representation of a domain only if each node X_i is conditionally independent of its predecessors given its parents. This property serves as a guideline for designing BNs: the parents of node X_i should contain those nodes in $X_1,\ldots X_{i-1}$ that *directly* influence X_i. Furthermore, it is also important to provide a correct node ordering in a way that nodes that directly influence other nodes are added first in the list, and then nodes they influence until the nodes which do not have any influence on the other nodes are reached. If the variables are ordered incorrectly, the resulting network structure may fail to reveal many conditional independencies among variables. In the final step of constructing a BN, conditional probabilities for each node are assessed. They represent a person's *degree of belief* regarding the occurrence of the encoded events in the domain[3]. For detail specification of the procedure how to design BNs from the prior knowledge about the domain, we refer to (Heckerman, 1995; Russell and Norvig, 2003; Jensen, 1996). In this chapter, we are concerned with techniques as to how to learn BNs from data including both the structure learning and conditional probability distributions, also referred to as parameters, learning. These techniques as well as the techniques for performing inference in BNs will be discussed in detail in following sections.

The central formula to probabilistic inference in BNs is Bayes' rule:

$$P(X|Y) = \frac{P(Y|X)\cdot P(X)}{P(Y)} \tag{4.4}$$

It is used to update probabilities of variables in BNs when new information is entered in the network by setting some of the variables to particular values.

Some of the statistical techniques for learning BNs from data are based on the Bayesian approach. In the Bayesian approach to learning the Bayes' rule is applied for combining

[3] The concept of probability defined as the degree of belief a person has in an uncertain proposition is specified in Bayesian probability theory

prior knowledge and data. The prior knowledge about possible hypotheses X ($P(X)$) is combined with information obtained from data Y given hypotheses X ($P(Y|X)$) to derive new posterior knowledge about hypotheses $P(X|Y)$.

In the common notation, $P(X|Y)$ is referred to as the *posterior*, $P(Y|X)$ is referred to as the *likelihood*, $P(X)$ is the *prior* and $P(Y)$ is the *evidence*.

4.2.1 Learning BNs from data

The problem of learning a BN from data can be stated as follows. Let $X = \{X_1, \ldots, X_n\}$ be a set of random variables. Given a training set $D = \{\mathbf{x^1}, \ldots, \mathbf{x^N}\}$ where each $\mathbf{x^i}$ is an instantiation of the variables in X, the task is to find a BN $B = (G, \theta)$ that best matches the data.

Learning a BN consists of two tasks: learning the BN structure and learning the conditional probability distributions (henceforth, parameters) given the structure. The complexity of both learning procedures increases with partial observability which refers to both having some unobservable or hidden variables in the model as well as having some observable variables which do not have values for all instances in the training set. In discussing the structure learning task, we assume full observability, that is, no missing data nor hidden variables. However, for parameters learning, we will present approaches regarding both complete and incomplete data sets. As in previous discussions, we focus on BNs over multivalued discrete random variables.

Structure learning

There are two general classes of approaches for learning the structure of a BN: the *CI-based*[4] approaches and *search&score* approaches. In CI-based approaches, the learning is seen as a constraint satisfaction problem. First, conditional independence properties are estimated among the variables on D using some statistical test such as the χ^2 or mutual information tests. Then, a network is designed consistent with identified dependencies and independencies among variables. An overview of some of the existing CI-based algorithms can be found in (Cheng et al., 1998). In search&score approaches, the learning is seen as an optimization problem. It consists of searching for a structure, through the space of possible structures, that maximizes an a priori defined scoring metric. The scoring metric describes the fitness of each possible structure to the data. As search&score approaches are widely used and they were also applied for the experiments presented in Chapter 5, we will discuss them in more detail.

Scoring metrics Scoring metrics, also referred to as quality measures, can be based on different approaches including a Bayesian approach (Cooper and Herskovits, 1992; Heckerman et al., 1995), a minimum length principle (MDL) approach (Lam and Bacchus, 1994) or an information criterion approach (Schwarz, 1978; Akaike, 1974). Further discussion on

[4] CI stands for conditional independence

score&search methods will be based on a Bayesian approach since it is used for the experiments presented in the Chapter 5. For a detailed discussion on the Bayesian approach for estimating the quality of a BN structure, we refer to (Cooper and Herskovits, 1992; Heckerman et al., 1995; Bouckaert, 1995).

The main idea of a Bayesian approach is the following: assuming that we are uncertain about the network structure that encodes the full joint probability distribution over the variables X, we encode this uncertainty by defining a variable G whose values correspond to the possible network structures. Then, we specify our degree of belief regarding the network structures with *prior* probability distribution P(G). When the training set D is available, the quality of the structure is estimated using the *posterior* probability of the network structure G given the data set D which is computed using the Bayes' rule as follows:

$$P(G|D) = \frac{P(D|G)P(G)}{P(D)}{}^{5} \tag{4.5}$$

Since $P(D)$ is equal for all network structures it can be treated as a constant and omitted from the estimation of quality measures. Therefore, it is sufficient to estimate the numerator in the right-hand side of the equation $P(G, D) = P(D|G)P(G)$. This probability can be computed as follows (Cooper and Herskovits, 1992):

$$P(G, D) = P(G) \int P(D|\theta, G)P(\theta|G)d\theta \tag{4.6}$$

As θ represents a set of parameters that are continuous variables, $P(\theta|G)$ denotes a conditional probability density function over θ given the network structure G. Since the integral in the equation (4.6) is calculated over all possible parameter assignments θ for the given structure G, it is actually estimated over all BNs with that structure. For practical reasons, the logarithm of the posterior probability is often used as a score function: $\log P(G, D) = logP(D|G) + logG$.

Several quality measures presented in the literature have been derived using the Bayesian approach. Cooper and Herskovits (1992) derived the following formula for estimating P(G,D) based on the assumption that no set of parameters is preferred for a given network structure, that is, the density function $P(\theta|G)$ is uniform:

$$P(G, D) = P(G) \cdot \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{(r_i - 1)!}{(N_{ij} + r_i - 1)!} \cdot \prod_{k=1}^{r_i} N_{ijk}! \tag{4.7}$$

where r_i denotes the cardinality of the value set of the node X_i $(1 \leq i \leq n)$, q_i is the cardinality of $Pa(X_i)$, that is, the number of different values to which parents of X_i can

[5] In order to compare the posterior probabilities of two network structures, the following equivalent criterion is often used:

$$\frac{P(G_1|D)}{P(G_2|D)} = \frac{P(D|G_1)P(G_1)}{P(D|G_2)P(G_2)}$$

be instantiated: $q_i = \prod_{X_j \in Pa(X_i)} r_j$. Assuming that values in the X_i value set as well as configurations of the $Pa(X_i)$ are ordered, N_{ijk} ($1 \leq i \leq n$, $1 \leq j \leq q_i$, $1 \leq k \leq r_i$) denotes the number of instances in D for which X_i takes its k-th value and $Pa(X_i)$ takes its j-th configuration; $N_{ij} = \sum_k N_{ijk}$ denotes the number of instances in D for which $Pa(X_i)$ takes its j-th configuration. The score metric defined with the equation (4.7) is referred to as the *K2 metric*.

Assuming that the density function, $P(\theta|G)$ is not uniform, but Dirichlet distributed, Heckerman et al. (1995) derived the *Bayesian metric*, also known as the *Bayesian Dirichlet (BD)* metric, that has the following form:

$$P(G,D) = P(G) \cdot \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(N'_{ij})}{\Gamma(N'_{ij} + N_{ij})} \cdot \prod_{k=1}^{r_i} \frac{\Gamma(N'_{ijk} + N_{ijk})}{\Gamma(N'_{ijk})} \tag{4.8}$$

where $\Gamma(.)$ is the gamma-function which satisfies the condition that $\Gamma(1) = 1$ and $\Gamma(n+1) = n\Gamma(n)$; N'_{ij} and N'_{ijk} represent choices of priors on counts that satisfy the condition that $\sum_{k=1}^{r_i} N'_{ijk} = N'_{ij}$. It is to be noted that the K2 metric is the special case of the BD metric where $N'_{ijk} = 1$ and consequently $N'_{ij} = r_i$. One disadvantage of the BD metric is that it assigns different score measures to likelihood equivalent networks (Heckerman et al., 1995): two network structures are considered as equivalent if they encode the same conditional independence relations between nodes (e.g. X⟶Y and X⟵Y); two equivalent network structures G_1 and G_2 are considered likelihood equivalent if $P(D|G_1) = P(D|G_2)$. Heckerman et al. (1995) proposed a likelihood-equivalent specialization of the BD metric called the *BDe* metric in which $N'_{ijk} = N/r_i \cdot q_i$ and $N'_{ij} = N/qi$ where N is the user's equivalent sample size, that is, the size of a data set that is equivalent to a user's knowledge.

The prior of the network structure P(G) as well as the priors on the parameters N'_{ijk} specify a users' knowledge about the domain. A structure learning process that makes use of these priors is thus based on the combination of the prior knowledge and data. The prior of the network structure can incorporate, for example, information about the 'real' network: a domain expert, for example, can suggest the existence of the specific edge with the specific orientation. Those network structures that closely resemble the real network will be given higher prior probabilities (Heckerman et al., 1995). If no prior information on network structure is available, P(G) is chosen to be uniform and thus can be omitted from the estimation of the quality of the network structure.

Search methods Having specified a quality measure, the goal is to find a network with the highest quality. In general, there may be more than one structure with the highest quality. For the purpose of simplicity, we assume that there is only one structure that maximizes the score metric. To find such a structure, one must search in the space of all possible structures. However, the number of possible structures grows exponentially with the number of variables, which indicates that it may not be feasible to find the structure with the highest quality in polynomial time. Moreover, it has been proven that finding such a structure is a NP-hard problem even when some restrictive assumptions are made such as that each node has at most $k > 1$ parents (Chickering, 1996; Chickering et al., 2004). Therefore, in past years, a significant amount of work in the machine learning community has been focused on developing heuristic-search methods for finding a good model.

Heuristics gain their computational efficiency by partially estimating the quality of each structure. This is made feasible due to *decomposability* of score metrics: a score metric is considered as decomposable if it can be represented as a product[6] of variable specific scores (Heckerman, 1995):

$$Q(G,D) = \prod_i q(X_i, Pa(X_i), D_i) \tag{4.9}$$

where D_i is the data restricted to the variables X_i and $Pa(X_i)$.

Due to decomposability each node can be considered independent of every other. This implies that instead of the complete network structure only a node and its parents need to be considered. Most commonly used heuristic search algorithms include greedy search with or without restarts, best-first search, simulated annealing and Monte Carlo search.

For the experiments presented in Chapter 5, we made use of the *K2 heuristic search algorithm* which is based on the greedy search strategy (Cooper and Herskovits, 1992). In the original version, the algorithm employs the K2 score metric for measuring the quality of a structure. However, other score metrics can be applied as well. The algorithm assumes that an ordering of the variables is given. This assumption reduces the number of possible alternatives since only predecessors of a variable are taken into consideration when determining its parents. The algorithm also assumes that an upper bound u on the number of parents that any node may have is specified. Given these assumptions, the K2 algorithm proceeds by making an assumption that a node has no parents and then adds incrementally the parent whose addition increases the most the probability of the given structure calculated by means of the q function in the equation (4.9). When the addition of no single parent can increase the probability of the given structure or the number of added nodes reaches the parent thresholds, the process of adding nodes to the parent set of the current node is terminated. The procedure is applied to all variables.

Parameter learning

Having designed the structure G of a BN from the prior knowledge about the domain, data or a combination of both, the next step in specifying the BN is the assignment of conditional probability distributions to each node in the network. In this section, we present methods for learning the conditional probabilities from the data with and without the use of prior knowledge.

Let $\boldsymbol{\Theta}_\mathbf{G}$ denote a multivalued variable whose values θ represent a set of real numbers which correspond to possible true values of the physical conditional probability distributions assigned to each node in the network G over the random variables $X = \{X_1, \ldots, X_n\}$: $\theta = (\theta_\mathbf{1}, \ldots, \theta_\mathbf{n})$. Similar to the structure learning, we assume that we have given a training set of N instances $D = \{\mathbf{x^{(1)}}, \ldots, \mathbf{x^{(N)}}\}$ where $\mathbf{x^{(i)}} = \{x_1^{(i)}, \ldots, x_n^{(i)}\}$.

[6] or a sum in a case where log is employed

where G denotes the given information that the full joint probability distribution over X is encoded in the network structure G.

In this section, we present two widely used approximation methods for assessing parameters when a complete data set is employed: the *Maximum Likelihood* (ML) and *Maximum A-Posteriori*(MAP) estimates. We also discuss the *Expectation Maximization* (EM) algorithm for estimating parameters for an incomplete data set.

Similar to the previous discussion, we assume that each variable $X_i \in X$ is a multinomial discrete variable having r_i possible values $x_i^1, \dots, x_i^{r_i}$. Furthermore, $Pa(X_i)$ is the parent set of the node X_i having q_i possible configurations $\mathbf{pa_i^1}, \dots, \mathbf{pa_i^{q_i}}$ where $q_i = \prod_{X_j \in Pa(X_i)} r_j$.

ML Estimation Seeing the data set D as a sample derived from an unknown true distribution, the ML estimator approximates this distribution by choosing the parameter set θ^* that maximizes the data likelihood:

$$\theta^* = \underset{\theta}{\operatorname{argmax}}\, \mathcal{L}(\theta) = \underset{\theta}{\operatorname{argmax}}\, P(D|\theta, G) \tag{4.10}$$

Assuming that given the true probability distribution represented with the BN $B = (G, \theta)$ the instances in the data set are independent, the likelihood in the equation (4.10) can be calculated as follows:

$$P(D|\theta, G) = \prod_{l=1}^{N} P(x^{(l)}|\theta, G) = \prod_{l=1}^{N} \prod_{i=1}^{n} P(x_i^{(l)}|\mathbf{pa_i^{(l)}}, \theta_i) \tag{4.11}$$

The formula on the right-hand side is a result of the estimation of the probability of each instance in the data set. This probability is estimated from the probability distribution factorized according to the structure G in terms of conditional probabilities that are specified by the parameter set θ.

A common approach in the ML estimation is to use the log likelihood instead of the likelihood:

$$\mathcal{LL}(\theta) = \log P(D|\theta, G) = \sum_{l=1}^{N} \sum_{i=1}^{n} \log P(x_i^{(l)}|\mathbf{pa_i^{(l)}}, \theta_i) = \sum_{i=1}^{n} \mathcal{LL}_i(\theta_i) \tag{4.12}$$

where $\mathcal{LL}_i(\theta_i) = \sum_{l=1}^{N} \log P(x_i^{(l)}|\mathbf{pa_i^{(l)}}, \theta_i)$. Each $\mathcal{LL}_i(\theta_i)$ can be maximized independently as a function of θ_i. In our case where X is a set of multinomial discrete random variables θ_i^* is a normalized conditional probability table containing counts of each value of X_i given each configuration of its parents $Pa(X_i)$: $P(X_i = x_i^k|Pa(X_i) = pa_j) = \frac{N_{ijk}}{N_{ij}}$.

The main problem when employing the ML estimator for learning parameters is that it is sensitive to the sparse data in a sense that there may be certain combinations of a variable's values and its parents' configurations that are not observed in a data ($N_{ijk} = 0$) which leads to the ML estimate of 0 for those combinations. The problem of defining some configurations as impossible because they are not observed on the training data can be prevented by adding the prior information to the network parameters. For that purpose, the MAP estimator can be employed.

MAP Estimation The MAP Estimation of the parameters of BNs is based on the Bayesian approach. The parameters θ are seen as an unknown variable governed by prior probability distribution $P(\theta|G)$ which represents some prior belief regarding the parameters given the information that the joint probability distribution over X can be encoded according to structure G. When the training set D is available, this belief is updated according to Bayes' rule as follows:

$$P(\theta|D,G) = \frac{P(D|\theta,G)P(\theta|G)}{P(D|G)} \tag{4.13}$$

The MAP estimate of θ is the expectation of θ with respect to our posterior belief about its values (Heckerman, 1995):

$$E_{P(\theta|D,G)}(\theta) = \int \theta P(\theta|D,G)d\theta \tag{4.14}$$

Since $P(D|G)$ in the equation (4.13) is not dependent on θ, it can be treated as a constant and $P(\theta|D,G)$ can be approximated as follows:

$$P(\theta|D,G) \approx P(D|\theta,G)P(\theta|G) \tag{4.15}$$

Let us first extend the notation introduced in this section. As previously defined θ is a vector containing a parameter set for each variable $\theta_{\mathbf{i}}$. Each $\theta_{\mathbf{i}}$ can be represented as a vector of parameter settings $(\theta_{\mathbf{i1}},\ldots,\theta_{\mathbf{iq_i}})$. Denoting $P(X_i = x_i^k|Pa(X_i) = pa_i^j) = \theta_{ijk}$, $\theta_{\mathbf{ij}}$ can be represented as the vector of parameters $(\theta_{ij1},\ldots,\theta_{ijr_i})$.

Assuming that the parameter vectors $\theta_{\mathbf{ij}}$ are mutually independent, the parameter prior can be estimated as follows:

$$P(\theta|G) = \prod_{i=1}^{n}\prod_{j=1}^{q_i} P(\theta_{\mathbf{ij}}|G) \tag{4.16}$$

Under the same assumptions the parameters remain independent given the data set D:

$$P(\theta|D,G) = \prod_{i=1}^{n}\prod_{j=1}^{q_i} P(\theta_{\mathbf{ij}}|D,G) \tag{4.17}$$

This indicates that each vector of parameters $\theta_{\mathbf{ij}}$ can be updated independently

$$P(\theta_{\mathbf{ij}}|D,G) \approx P(D|\theta_{\mathbf{ij}},G)P(\theta_{\mathbf{ij}}|G) \tag{4.18}$$

A common approach to modeling prior belief over multinomial variables is to use the Dirichlet distribution which has a property that the posterior distribution belongs to the same conjugate family as the prior distribution. For each vector of parameters θ_{ij}, the Dirichlet prior distribution is defined as follows:

$$P(\theta_{ij}|G) = Dir(\theta_{ij}|\alpha_{ij1},\ldots,\alpha_{ijr_i}) = \frac{\Gamma(\alpha_{ij})}{\Gamma(\prod_{k=1}^{r_i}\alpha_{ijk})}\prod_{k=1}^{r_i}(\theta_{ijk})^{\alpha_{ijk}-1} \tag{4.19}$$

The parameters α_{ijk} are often called hyperparameters in order to distinguish them from the parameters of the multinomial distribution θ_{ijk}. The parameters α_{ijk} can be seen as 'prior observation counts' for events governed by θ_{ijk}: the number of times that an expert has previously observed the instantiation of $X_i = x_i^k$ and $Pa(X_i) = \mathbf{pa_i^j}$. As previously mentioned, the posterior probability of the parameters $\theta_{\mathbf{ij}}$ is also Dirichlet and can be estimated using the equation (4.18)

$$P(\theta_{\mathbf{ij}}|D, G) \approx \prod_{k=1}^{r_i} (\theta_{ijk})^{N_{ijk}} \prod_{k=1}^{r_i} (\theta_{ijk})^{\alpha_{ijk}-1} \approx \prod_{k=1}^{r_i} (\theta_{ijk})^{\alpha_{ijk}+N_{ijk}-1} \tag{4.20}$$

Therefore, the posterior distribution $P(\theta_{\mathbf{ij}}|D, G)$ is the Dirichlet distribution $Dir(\theta_{\mathbf{ij}}|\alpha_{ij1}+ N_{ij1}, \ldots, \alpha_{ijr_i} + N_{ijr_i})$

Under the assumption of the parameter independence, the expectation of the parameters in respect to the posterior distribution as defined in (4.14) can be estimated as follows:

$$\theta^*_{ijk} = \frac{\alpha_{ijk} + N_{ijk}}{\alpha_{ij} + N_{ij}} \tag{4.21}$$

where $\alpha_{ij} = \sum_{k=1}^{r_i} \alpha_{ijk}$ and $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$. For detailed derivation of this formula, we refer to (Heckerman, 1995). In practice however, it is difficult to specify hyperparameters α_{ijk} for each combination of a variable's values and its parents configuration $(x_i^k, \mathbf{pa}_i^j)$. Several types of Dirichlet priors widely used in specifying network parameters, although less informative, have been described previously when different variants of the BD score metric were introduced.

EM The EM algorithm is the most well-known and widely used algorithm for learning parameters of BNs from incomplete data (Dempster et al., 1977). It is an iterative algorithm where each iteration is performed in two steps: the expectation (E) step followed by the (M) maximization step. In its original form, the algorithm converges towards the ML parameter estimates although it can also be used for finding MAP estimates by performing the MAP estimation in the M step instead of the ML estimation.

Let Y denote incomplete data consisting of the values of observable variables and Z denote either missing data or hidden variables. Y and Z form together the complete data set $D = \{Y, Z\}$. Since it is not feasible to estimate the complete data set log likelihood required for the ML estimate (see 4.10), we can calculate the expected log likelihood of the complete data set using our state of the knowledge regarding unobservable variables Z given by the posterior distribution $P(Z|Y, \theta)$. This calculation corresponds to the E step of the EM algorithm. In the subsequent M step, the ML estimates of the parameters are performed by maximizing the expected likelihood found on the E step. The estimated parameters in the M step are then used in the subsequent E step.

A more formal and comprehended specification of the EM algorithm can be found in (Bishop, 2006). Given the joint distribution over Y and Z governed by parameters θ, the goal of the EM estimate is to maximize likelihood function $P(Y|\theta)$ given by:

$$P(Y|\theta) = \sum_Z P(Y, Z|\theta)$$

or equivalently:

$$\log P(Y|\theta) = \log\{\sum_Z P(Y, Z|\theta)\}$$

The EM algorithm, as described in (Bishop, 2006), consists of the following steps

1. Choose an initial set of parameters θ^{old} in some way.
2. E step: Calculate $P(Z|Y, \theta^{old})$.
3. M step: Estimate θ^{new} by maximizing conditional expectation of the complete data log-likelihood with respect the posterior probability distribution found in the E step

$$\theta^{new} = \operatorname*{argmax}_{\theta} Q(\theta, \theta^{old}) = \operatorname*{argmax}_{\theta} \sum_Z P(Z|Y, \theta^{old}) \log P(Y, Z|\theta) \tag{4.22}$$

4. If neither the convergence criterion of the log-likelihood nor of the parameter values is satisfied then let $\theta^{new} = \theta^{old}$ and repeat the step 2.

The E step can be seen as representing the unknown value of Z by a distribution of values, and the M step as performing the ML estimation for the joint data obtained by combining this with the known value of Z. As previously mentioned, the EM algorithm can be used for the MAP estimation instead of the ML estimation by applying the MAP estimates in the M step, that is, by maximizing $Q(\theta, \theta^{\mathbf{old}}) + \log P(\theta)$.

4.2.2 Inference in BNs

One of the main applications of probabilistic models in general and thus of BNs in particular is the derivation of the probabilities of interest regarding the domain being modeled. A computation of probabilities of interest in a probabilistic model is known as a probabilistic inference. Since a BN denotes a representation of the full joint probability distribution over all variables in the domain $P(X_1, \ldots, X_n)$, it contains sufficient information for computing the probabilities of interest.

In a general form, inference can be specified in the following way. Let $E \subset X$ denote a set of variables for which evidence e has been obtained, or in terms of a graphical representation, a set of nodes whose values are observed. The probability distribution of a set of variables $Y \subset X$ given the observed evidence e can be computed from the full joint probability distribution as follows:

$$P(Y|e) = \frac{P(Y, e)}{P(e)} = \frac{\sum_z P(Y, z, e)}{\sum_{y,z} P(y, z, e)} \tag{4.23}$$

where z denotes values of all other variables in X, excluding Y and E. Variables in Y are often referred to as *query variables.*

Although BNs determine a compact representation of the joint probability distribution which enables a more efficient computation of the numerator and the denominator in the equation (4.23), the inference in BNs has been shown to be a NP-hard problem (Cooper, 1990b). Moreover, it has also been proven that approximate probabilistic inference in BNs is NP-hard (Dagum and Luby, 1993). Therefore, it is unfeasible to develop an exact or an approximate inference algorithm that can be applied to all classes of BNs.

As to the exact inference, the simplest approach that sums out all non-query variables from the full joint probability distribution factorized in terms of a BN is known as *inference by enumeration*[7]. A more efficient approach based on a similar idea is the *variable elimination* technique. It is based on the interleave of sums and products in a way which enables that repeated calculations are cached and saved for future use. Furthermore, a variable elimination process may also include pruning of the irrelevant variables as it was found that every variable that is not an ancestor of a query variable or evidence variable in the network is irrelevant for the query (Zhang and Poole, 1994; Dechter, 1996).

Instead of viewing inference as an elimination process, we can see inference as a message-passing process where messages are passed locally among surrounding nodes. Pearl (1982) introduced the *message-passing*[8] algorithm applicable to a simple class of graphs called *trees*: graphs in which there is one and only one undirected path between any two nodes in the graph. Receiving a message from its children, a node sends a message, that is, an updated probability distribution to its parent and other children. Similarly, an information received from the node's parent is updated and sent to its children. For each edge in a network, the message is passed only once in each direction: from a parent to a child and vice versa. The algorithm proceeds by entering evidence on the evidence variables and passing messages locally through the network. The message passing algorithm has been extended by (Pearl, 1986, 1988) for *single-connected networks*, also known as *polytrees*: graphs in which there is at most one undirected path between any two nodes in the network.

However, the most problematic class of graphs from the inference perspective are graphs with undirected cycles - *multiple-connected networks.* There are two main approaches developed for the exact inference in the multiple-connected graphs: junction tree propagation approach and cutset conditioning approach. The *junction tree propagation* algorithm (Lauritzen and Spiegelhalter, 1988; Jensen et al., 1990) is based on transforming a multiple-connected network into a probabilistic equivalent tree - a so-called junction tree - where each node in a tree contains a subset of the variables from the original set. Then, the message-passing algorithm is applied on the created junction tree for deriving conditional and marginal probabilities of interests. The cutset conditioning algorithm (Pearl, 1986), on the other hand, transforms a multiple-connected network into a set of single connected networks by instantiating variables to the definite values. The probability of interests, conditional or marginal, are computed as a weighted average over the values computed by

[7] the term taken from (Russell and Norvig, 2003)

[8] also known as the *belief propagation* or the *sum-product* algorithm

each polytree. Additional approaches to exact inference proposed in the literature comprise *recursive decomposition* (Cooper, 1990a) and *symbolic inference* (Shachter et al., 1990).

In many situations, however, exact inference may be intractable. This could be, for example, because of the large number of hidden variables included in a model or because of the complex form of the posterior probability distribution, in a case of the continuous variables, which prohibit analytical or numerical computation being applied. In such situations, the approximate inference can be applied.

Approximate approaches can be classified as *deterministic* or as *stochastic.* Deterministic approaches include *variational inference* techniques which are concerned with approximating true posterior probability distribution with a simpler variational distribution which is chosen in a way to be tractable and to converge to the posterior in reasonable time. For an overview of the existing variational inference techniques, we refer to (Jordan et al., 1999; Bishop, 2006). Minka (2001) proposed another deterministic approximate inference known as *expectation propagation.*

Stochastic approaches are based on numerical sampling techniques also called Monte Carlo algorithms. Samples are generated using BN either from conditional or from unconditional probabilities. The sampling approaches can be categorized as direct sampling or as Markov chain sampling approaches. The direct sampling approach comprises rejection sampling, importance sampling and likelihood weighting. The *rejection sampling* first generates samples from a BN in the topological order without consideration of the evidence and then rejects those samples that do not match evidence. For the discrete variables, $P(X_i = x_i|e)$ is estimated by counting how often the event $X_i = x_i$ occurs in the remaining samples. The main disadvantage of this method is that it may reject a large number of samples, especially if the evidence contains a large number of variables. *Importance sampling*, on the other hand, does not reject samples but generates samples consistent with the evidence weighted with importance weights. Likelihood weighting is an improvement of importance sampling in which all evidence variables are instantiated to their values and all other variables X_i are sampled from $P(X_i|Pa(X_i))$ in which the variables $Pa(X_i)$ are set to their sampled values. Each event is weighted by the likelihood that the event accord to the evidence which is calculated as the product of conditional probability distribution of each evidence variable given its parents.

Another group of sampling methods are Markov Chain Monte Carlo (MCMC) methods. In contrast to previously mentioned approaches, these methods generate each sample by making a random change in the previously generated sample by changing a value for one of the query variables condition on the current sampled values of the variables in its Markov blanket. In this way, a sequence of generated samples forms a Markov chain in a sense that the probability distribution of the next event is conditionally dependent only on the current event given all previously generated events. Detailed descriptions of different MCMC algorithms can be found in(Bishop, 2006).

4.3 BNs as Classifiers

In the previous sections, we described BNs as powerful tools for knowledge representation and inference in the domains that deal with uncertainty. In this section, we discuss their usage for the classification task.

Using the learning methods described in the previous section, a BN which encodes the full joint probability distribution $P(A_1, \ldots, A_n, C)$ can be induced from the data and then used as a classifier by inferring posterior probability distribution of the class variable C given the values $\{a_1, \ldots, a_n\}$ of attributes $\{A_1, \ldots, A_n\}$: $P(C|a_1, \ldots, a_n)$. This procedure is often referred to in the literature as *unsupervised* in the sense that the learning procedure does not distinguish the class variable from the attributes. In the decision stage of classification, the class with the highest posterior probability is assigned to $\{a_1, \ldots, a_n\}$.

BNs had not been considered as classification tools until (Langley et al., 1992) found that simple BNs, called Naive Bayes (NB), which assume that all attribute nodes are conditionally independent given the class variable, show good classification performances in comparison to performances of some other classification algorithms. Friedman et al. (1997) have shown that unrestricted BNs, which relax the strict and in many cases unrealistic requirements for conditional independence encoded in NB networks, in a large number of cases fail to outperform NB classifiers when used in an unsupervised way. The reason for poor classification performances of unrestricted BNs, as theoretically explained by Friedman et al. (1997) and Greiner et al. (1997), lies in the mismatch between functions that score-based learning algorithms, on the one hand, and classification algorithms, on the other hand, aim to maximize. The learning algorithms maximize scoring functions expressed in terms of the data likelihood which specifies how well a BN structure fit the data whereas the classification algorithms maximize prediction accuracies expressed in terms of the conditional likelihood.

As shown in the equation (4.5), a score metric contains a term that measures likelihood of the structure G given the data set D which can be written as follows:

$$L(G|D) = P(D|G) = \prod_{i=1}^{N} P(c^i|a_1^i, \ldots, a_n^i, G)P(a_1^i, \ldots, a_n^i|G) \tag{4.24}$$

or in terms of log-likelihood:

$$L(G|D) = P(D|G) = \sum_{i=1}^{N} \log P(c^i|a_1^i, \ldots, a_n^i, G) + \sum_{i=1}^{N} \log P(a_1^i, \ldots, a_n^i|G) \tag{4.25}$$

where each instance $\mathbf{x}^i$ in the data set $D = \{\mathbf{x}^1, \ldots, \mathbf{x}^N\}$ contains assigned values to the attributes $A_1, \ldots, A_n$ and to the class variable C: $\{a_1^i, \ldots, a_n^i, c^i\}$

The first term in the equation 4.25 measures how well the network G estimates $P(C|A_1, \ldots, A_n)$. It is thus related to the score of the network as a classifier. The second term represents the joint probability distribution over the attributes estimated according to network G.

Friedman et al. (1997) have shown that the first term is dominated by the second term when there are many attributes. This implies that relatively large error in the first term may not be reflected in the total score function. Comparing unrestricted BN classifiers with the NB classifiers over various data sets, Friedman et al. (1997) found that cases in which NB classifiers outperform BN classifiers although having the worse total scores, are cases in which the Markov blanket of the class variable contains a small number of attributes. From all these, it follows that a network with a better score is not necessarily a better classifier.

One solution to this problem proposed by (Friedman et al., 1997) is to restrict the score metric only to the first term estimating in this way *conditional-log likelihood*:

$$CLL(G|D) = \sum_{i=1}^{N} \log P(c^i|a_1^i, \dots, a_n^i)$$

However, this score, as opposed to the log-likelihood, is not decomposable over the network structure (see 4.9). Since it is not feasible to maximize the choice of each conditional probability distribution in the network separately, the computation of the network that maximize $CLL(G|D)$ becomes unfeasible. Therefore, Friedman et al. (1997) abandoned this solution leaving as an open issue the search for some heuristic approaches that will allow effective learning of BN classifiers by maximizing the conditional log likelihood score. Recently, the research on developing more accurate BN classifiers has begun to develop in this direction (Grossman and Domingos, 2004; Jing et al., 2005).

Friedman et al. (1997) proposed another approach in learning BN classifiers that is concerned with augmenting the NB structure with edges among attributes. It ensures that all attributes are in the Markov blanket of the class node. Following this approach several types of BN classifiers have been developed among which are the Tree Augmented Naive Bayes (TAN) classifier and the Bayesian Augmented Naive Bayes (BAN) classifiers. Those classifiers in addition to the NB and general Bayesian Network (GBN) classifiers were employed for the experiments presented in Chapter 5. All four types of classifiers are presented in more detail in the following sections. These classifiers differ based on the structures that are permitted. Once the structure is known, any of the learning algorithms presented in Section 4.2.1 can be applied for estimating the network parameters. Therefore, in the following sections, we will discuss only methods for learning the structure of each classifier.

Naive Bayes (NB)

A NB classifier is a simple BN which encodes the class variable as the parent of all attribute variables. It also imposes a strict conditional independence among the attribute variables given the class prohibiting in that way any connections between the attribute nodes in the network (see Figure 4.2). Since the structure is known in advance, learning NBs from data is concerned only with the parameter learning. Due to the strict conditional independence requirement, the posterior distribution of the class given the attributes is

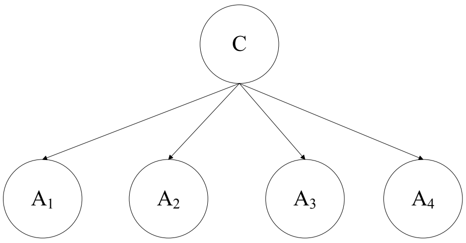


Figure 4.2: An example of a NB structure

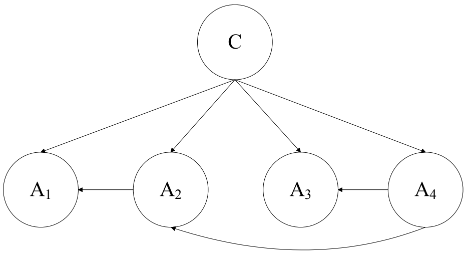


Figure 4.3: An example of a TAN structure

estimated in the following way:

$$P(C|a_1, \ldots, a_n) \approx P(C) \prod_{i=1}^{n} P(a_i|C)$$

Although the independence criterion among attributes is unrealistic in many domains, NB classifiers have shown surprisingly good performances in comparison to more sophisticated classifiers over a large number of data sets especially in those cases where attributes are not strongly correlated. For a detailed technical explanation of the reasons for the surprisingly good performances of NBs classifiers we refer to (Friedman, 1997a).

Tree Augmented Naive Bayes (TAN)

A TAN classifier is an extension of a NB classifier which allows the attribute nodes to form a tree. The algorithm for learning TANs introduced by Friedman et al. (1997), first learns a tree structure over $D \setminus \{C\}$ by applying conditional mutual information tests $I(A_i; A_j|C)$ defined as:

$$I_P(A_i; A_j|C) = \sum_{a_i, a_j, c} P(a_i, a_j, c) \log \frac{P(a_i, a_j|c)}{P(a_i|c)P(a_j|c)} \tag{4.26}$$

This function measures how much information A_j provides about A_i when C is known. The tree is created in three steps as stated in (Friedman et al., 1997). First, it builds a complete undirected graph in which vertices are attribute variables. An edge that connects nodes A_i and A_j is annotated with the weight $I_P(A_i; A_j|C)$. Second, the algorithm forms a maximum weighted spanning tree. Finally, it transforms the created undirected tree to a direct one. After the tree has been created, the class node is added as a parent to each attribute node forming in this way a structure similar to the NB structure (see Figure 4.3). It is to be noted that this algorithm can be classified as a *supervised* CI-based structure learning approach (see Section 4.2.1) - supervised in the sense that the class node is treated differently than the attribute nodes.

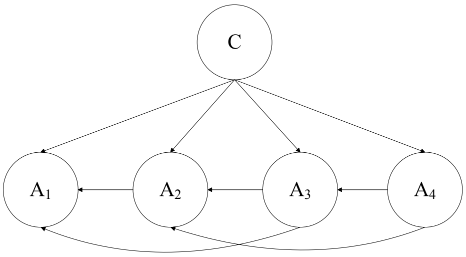


Figure 4.4: An example of a BAN structure

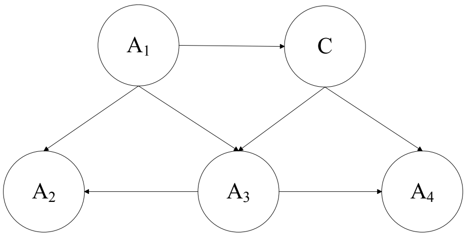


Figure 4.5: An example of a GBN structure

Bayesian Augmented Naive Bayes (BAN)

BAN classifiers extend NB classifiers by allowing attributes to create an arbitrary graph (see Figure 4.4). The algorithm for learning BAN classifiers is similar to the TAN learning algorithm. The only difference is that a BAN learner applies an unrestricted BN structure learning algorithm instead of tree learning algorithm. Both CI-based and score-based methods can be applied for learning BNs over the attributes variables (Friedman et al., 1997; Cheng and Greiner, 1999).

General Bayesian Network (GBN)

In contrast to the NB and augmented NB classifiers, GBN classifiers treat the class node as an ordinary node applying in that way an unsupervised approach for learning the network structure over the class and attribute variables (see Figure 4.5). The GBN learner finds the Markov blanket of the class node performing in that way the attribute selection.

4.4 Dynamic Bayesian Networks

In the previous section, we presented BNs as probabilistic models that assume the static environment, that is time-invariant distribution and time-invariant data generation models. In this section, we discuss Dynamic Bayesian Networks (DBNs) which represent general probabilistic models of sequential data. Sequential data may be either time-series data generated by a dynamic system or sequence data. The term “dynamic” is used to denote that a dynamic system is being modeled and not that the network is being changed over time. DBNs can be seen as an extension of BNs especially aimed at probabilistic sequential modeling.

DBNs belong to a family of so-called *state-space models* which assume that there is some underlying hidden state of the world which generates observations (or outputs) and that hidden state evolves over time (Murphy, 2002). Additionally, the evolution of a hidden state can be determined as a function of some input variables. The two most common state-space models are Hidden Markov Models (HMMs) and Kalman Filter Models (KFMs). DBNs are a more general and expressive framework for representing state-space models. They generalize HMMs and KFMs by allowing complex interdependencies among variables in the networks; HMMs and KFMs have a restricted topology and they can be seen as simple forms of DBNs. The difference between a DBN and a HMM is that the hidden state in a DBN is represented in terms of a set of random variables whereas in a HMM it is modeled as a single discrete random variable. DBNs and KFMs differ regarding the type of nodes allowed in the network: all nodes in KFMs are required to be linear-Gaussian whereas DBNs do not impose any restriction on the node types.

4.4.1 Representation

Let $\mathcal{Z} = Z_1, Z_2, \ldots$ denote semi-infinitive collections of random variables each of which denoting a random variable at particular time t, represented as $Z_t = (X_t, Y_t)$ where X_t represents a set of hidden variables, often referred to as *states*, and Y_t represents a set of observational variables. Additionally, Z_t may also include the control variables U_t used for modeling controlled dynamic systems. As controlled variables are not considered in this thesis, we omit U_t from further discussion in order to simplify the notation. Furthermore, for the same reasons, we will assume that a DBN is a first-order Markov model, that is, the state of the model at the time t is dependent only on the state of the model at time *t-1*. However, a DBN can also be specified as a higher-order Markov model. As stated in (Murphy, 2002), a DBN over $\mathcal{Z}$ can be formally defined as follows:

4.4.1. Definition. A DBN over semi-infinite collections of random variables $Z_1, Z_2, \ldots$ is a pair $(B_1, B_{\longrightarrow})$ where B_1 is a *prior BN* that specifies a prior distribution over the initial set of states and observations $P(Z_1)$ and B$_{\longrightarrow}$ is a two-slice temporal BN (2TBN), so-called *transition network*, which specifies $P(Z_t|Z_{t-1})$ by means of a directed acyclic graph as follows:

$$P(Z_t|Z_{t-1}) = \prod_{i=1}^{n} P(Z_t^i|Pa(Z_t^i)) \tag{4.27}$$

where Z_t^i represent *i-th* node at time t which can be either from X_t or from Y_t. $Pa(Z_t^i)$ is a set of parents of the node Z_t^i which can be either from the same time-slice or from the previous time-slice. The nodes in the first slice of a 2TBN do not have any parameters assigned to them. However, each node in the second slice of the DBN has an associated conditional probability distribution specified as $P(Z_t^i|Pa(Z_t^i))$ for $t > 1$.

The provided definition covers the general case in which the prior network may have a quite different independence structure than the transition network. However, very often a DBN is represented using only the transition network assuming that all slices have the same structure. An example of such a network is given in Figure 5.3. Connections between variables within a time slice are referred to as *intra-slice* connections whereas connections between time slices are known as *inter-slice* connections.

In general, it can be assumed that parameters are time-invariant, that is, $P(Z_t^i|Pa(Z_t^i)) = P(Z_{t+l}^i|Pa(Z_{t+l}^i))$. If parameters do change over time they can be added to a model in terms

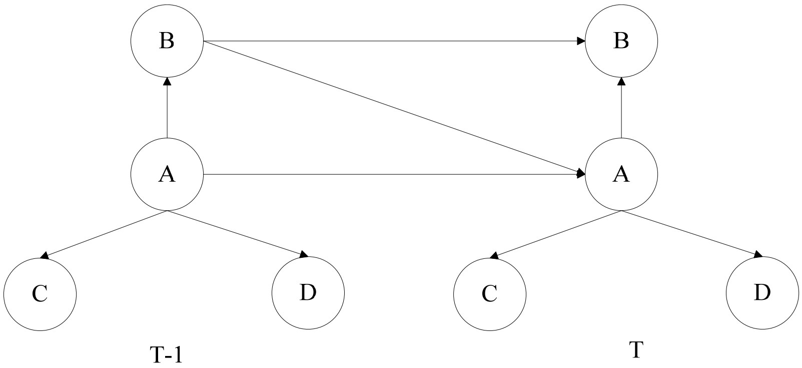


Figure 4.6: An example of a DBN structure

of random variables. Since parameters are tied across time slices, it is possible to model sequences of unbounded length by unrolling the 2TBN into a BN structure. Given a DBN model, the full joint probability distribution over a set of variables that form a sequence of length T is represented as follows:

$$P(Z_{1:T}) = \prod_{t=1}^{T}\prod_{i=1}^{n} P(Z_t^i|Pa(Z_t^i)) \tag{4.28}$$

4.4.2 Learning DBNs from data

The techniques for learning DBNs from the data are mostly straightforward extensions of the techniques for learning BNs described in Section 4.2.1. In contrast to the static case, the training set D consists of N_s observational sequences. The *k-th* such sequence having the length N_k specifies values $\{z_1^k, \dots, z_{N_k}^k\}$ for the variables $Z_1^k, \dots, Z_{N_k}^k$. Such a data set provides N_s instances of initial slices Z_1^i used for training the prior network B_1 and $N = \sum_k N_k$ instances of transitions from which $B_{\longrightarrow}$ can be estimated. If there are neither hidden states X_t involved in the model nor missing observations for observational variable Y_t^k, the data set D is considered as complete.

Structure learning

Friedman et al. (1998) proposed methods for learning structures of DBNs from complete as well as from incomplete data designed as an extension of the techniques developed for the static case. Regarding the complete data set, Friedman et al. (1998) defined extensions of the BDe criterion as well as of an information criterion known as BIC (Bayesian Information Criterion) for the application to DBNs (see Section 4.2.1). Similar to static BNs, a greedy search can be employed for searching for a structure that maximizes the specified score metrics. An additional constraint is that the network structure needs to be repeated over time. For learning the structure from an incomplete data set, Friedman et al. (1998) proposed an extension of the structural EM algorithm (SEM), originally developed for static BNs by Friedman (1997b), for DBNs. The SEM algorithm is a modification of the EM algorithm which in the M-step performs not only re-estimation of the network parameters but also uses the expected counts according to the current structure to evaluate any other candidate structure. In addition to the complete structure learning which comprises both learning the intra-slice and inter-slice connectivity, it is also common to specify intra-slice connectivity in advance and to reduce the learning problem to learning connections across the slices.

Parameter learning

Given the structure of a DBN, learning the network parameters can be performed offline and online. In this thesis, we made use of the offline learning approach. Similar to learning parameters of static BNs, the offline learning assumes that the data set D containing N_s sequences of the same or different lengths is given in advance. The same techniques as those used for parameter learning in the static case can be applied in the dynamic case: the ML or MAP estimator if D is complete; otherwise, the EM estimator. The parameters of the prior network are often taken to represent the initial state of the dynamic system being modelled. In this case, these parameters are estimated independently of the parameters for the transition network. Very often, especially when N_s is very small, these parameters are fixed a priori. However, if the parameters represent stationary distribution of the system, then their estimation is coupled with estimation of the parameters of the transition network.

4.4.3 Inference in DBNs

The general inference problem in DBNs it to compute the probability distribution $P(X_t^i|y_{1:t_1})$ where X_t^i denotes the *i-th* unobserved variable at slice t and $y_{1:t_1}$ is a set of t_1 consecutive observations, that is, evidence between starting time and t_1. The general inference problem can be refined into the following specific inference tasks.

- Filtering: $P(X_t|y_{1:t})$
- Smoothing: fixed-leg smoothing $P(X_{t-l}|y_{1:t})$ and fixed-interval or offline smoothing $P(X_t|y_{1:T})$, where T is the number of the slices in the observed sequence.
- Prediction: $P(X_{t+l}|y_{1:t})$
- Viterbi decoding: $x^*_{1:t} = \text{argmax}_{x_{1:t}} P(x_{1:t}|y_{1:t})$

The ways these types of inferences are calculated are dependent on the models and algorithms used. The term model at this instance refers to the general DBNs discussed so far as well as to different types of specialized DBNs which have a specific network structure and an assigned set of properties.

Similar to static BNs both exact and approximate inference methods can be applied. A common approach to exact inference in a DBN is to unroll the DBN for T slices and thus apply any inference algorithm to the resulting static BN. Murphy (2002) presented the application of the junction tree algorithm to the unrolled DBN for the offline smoothing. If the DBN contains discrete state nodes, it can be first converted to a HMM and then the well-known forwards-backwards (FB) algorithm for smoothing in HMM can be applied. To estimate $P(X_t = i|y_{1:T})$[9] the FB algorithm combines $\alpha_t(i) = P(X_t = i|y_{1:t})$ computed in the forward pass and $\beta_t(i) = P(y_{t+1,T}|X_t = i)$ computed in the backward pass. In this thesis, we made use of the junction tree algorithm.

As to the approximate inference in DBNs, both deterministic and sampling approaches as defined in Section 4.2.2 can be applied. Standard deterministic approaches for approximate inference for discrete DBNs are the Boyen-Koller algorithm (Boyen and Koller, 1998) and the Factored Frontier algorithm (Murphy and Weiss, 2001). Examples of the sampling approaches include the importance sampling and MCMC methods (see Section 4.2.2). For a detailed overview of the existing approximate inference algorithms, we refer to (Murphy, 2002).

4.5 Classifier evaluation

Once a classifier has been learned using a training data set, we need to estimate how predictive the classifier is, that is, how well the classifier performs on new unseen data. In this section, we discuss different metrics and methods used for the evaluation of classifiers' performances.

4.5.1 Evaluation metrics

There are several criteria that can be used in assessing a classifier's performance such as success rate, risk of making incorrect predictions, accuracy of probability estimates, interpretability, time and special complexity regarding both training and testing (Witten and Frank, 2000). Which of these criteria are to be used depends on the particular application of the classifier.

A natural measure for assessing the overall performance of a classifier is the *accuracy*, also known as *success rate* defined as the proportion of correctly classified instances in a data set[10]. However, it has been shown in the literature that overall accuracy may not be an accurate estimator of a classifier's performances if class values are unequally distributed in the data set. Therefore, a number of different performance measures has been proposed in the literature for a more reliable assessment of classifiers' qualities. For a detailed overview of some of existing standard evaluation metrics, we refer to (Witten and Frank, 2000). In this thesis, we made use of the *precision*, *recall* and *F-measure* for estimating how well

[9] Since HMMs contain one discrete hidden node per slice the event $X_t = i$ denotes that X_t takes the *i-th* value. A DBN, on the other hand, may have several hidden nodes and so far, X_t^i has been used to denote the i-th node in the slice t. However, in this particular case, the DBN is converted to HMM. Therefore, the HMM notation is employed

[10] alternatively, the *error rate* is used defined as $100\% - S_r$ where S_r is the success rate

classifiers perform regarding the prediction of a particular class value:

$$\text{Recall}_{\text{C}_i} = \frac{\text{number of correctly classified C}_i}{\text{total number of C}_i \text{ in data}} \tag{4.29}$$

$$\text{Precision}_{\text{C}_i} = \frac{\text{number of correctly classified C}_i}{\text{total number instances classified as C}_i} \tag{4.30}$$

$$\text{F-measure}_{\text{C}_i} = \frac{2 \times \text{Recall}_{\text{C}_i} \times \text{Precision}_{\text{C}_i}}{\text{Recall}_{\text{C}_i} + \text{Precision}_{\text{C}_i}} \tag{4.31}$$

4.5.2 Evaluation methodology

To predict a classifier's performance on unseen data, performance measures need to be estimated on a new data set, referred to as a *test set*, which contains instances that were not used in the learning process. When a large amount of data is available, the common practice is to divide the data set into a training set and a test set where the training set forms a larger data partition; usually, two-thirds of the data are used for training and one-third for testing. However, when the amount of data for training and testing is limited, it is desirable to apply the cross-validation technique. Cross-validation assumes that the data set is divided into n approximately equal data partitions, so-called *folds*, each of which is used for testing the classifier trained on the remaining $n - 1$ folds. This process is called *n-fold cross validation.* Repeating the evaluation process n times using non-overlapping test sets reduces the bias caused by the choice of the instances included in the training and test sets. The overall error estimate is computed as an average of the error estimates obtained on each fold.

An important issue when splitting the data set into a training and test sets or into n folds, is to ensure that class values are almost equally distributed in the training and test set. This process, known as *stratification*, reduces the variance of the estimated measures. As discussed in (Witten and Frank, 2000), empirical tests on various data sets using a wide range of classifiers have shown that the most accurate estimates of classifiers performances are obtained using *stratified 10-fold cross validation.* In some cases, an additional data set, a so-called *validation set*, can be employed for the optimization of parameters of a classifier.

Chapter 5

Addressee classification in meetings using Bayesian Networks

In this chapter, we present results on addressee classification in four-participants face-to-face meetings using several types of static BN classifiers as well as using DBN classifiers.

The experiments presented in this chapter were conducted on the M4 and the AMI meeting corpora described in Chapter 3. A part of the M4 corpus that consists of the annotated M4 meetings is used in our experiments. As discussed in Chapter 3, the M4 meetings are discussion meetings that are scripted in terms of type and schedule of meeting activities, but the content is natural and unconstrained. The AMI corpus, on the other hand, is a large collection of (1) scenario-based meetings involving a design team focused on the development of a TV remote control prototype and (2) naturally occurring meetings in a wide range of domains. A part of the AMI scenario-based collection that was annotated with addressee as well as with phenomena relevant to addressing was employed in our experiments. The M4 and most of the selected AMI meetings were recorded at the IDIAP smart meeting room; only two AMI meetings recorded in the TNO and Edinburgh meeting rooms were annotated with addressee information and thus exploited in our experiments. Although the layout of the rooms is similar for both collections, in the AMI scenario, the rooms were equipped with additional "attention distracters" such as the task object, in the first place, and laptops. Furthermore, participants in the AMI meetings, in contrast to the M4 meetings, were assigned the roles of project manager (PM), industrial designer (ID), user interface designer (UI) and marketing expert (ME) with different types and degrees of responsibilities concerning the design project as the whole.

The experiments presented in this chapter were conducted using five sorts of features: utterance features, gaze features, conversational context features, meeting context features and participant roles features. The selection of the features was motivated by the analysis of addressing mechanisms presented Chapter 2. Although it was shown that deictic hand and head gestures are used as means of addressing, instances of this type were found to occur very rarely in the data.

First, we conducted experiments on the M4 data using NB and BAN classifiers. The experiments should be seen as preliminary explorations of appropriate features and models for addressee identification in face-to-face meetings. The results of the experiments are reported in Section 5.3. Second, we performed experiments on the AMI data using several static BN classifiers and the DBN classifier. The results are presented in Section 5.4. In addition to the NB and BAN classifiers, we evaluated performances of the TAN and GBN classifiers for the task of addressee prediction. These static BN classifiers were trained and tested on the AMI data, first, using the set of features employed for the experiments on the M4 data (henceforth, the M4 feature set) (Section 5.4.2). Then, they were evaluated using a modified set of contextual, gaze, utterance, meeting context and role features referred to as the AMI feature set (Section 5.4.3). Finally, as the contextual feature set includes information about the addressee of the immediately preceding dialogue act, we investigated how well the addressee of the current dialogue act can be predicted using the classified instead of hand labeled value for the addressee of the previous dialogue act. For that purpose, the DBN classifier was employed. The evaluation of the performances of the DBN classifier on the AMI data is presented in Section 5.4.4. The overall experimental setup is summarized in Figure 5.1. At the end of this chapter, we address the issue of further automation of the addressee classification using automatically extracted instead of manually annotated features (Section 5.5). Concluding remarks regarding the presented results as well as recommendations for future research on addressee identification are given in Section 5.6.

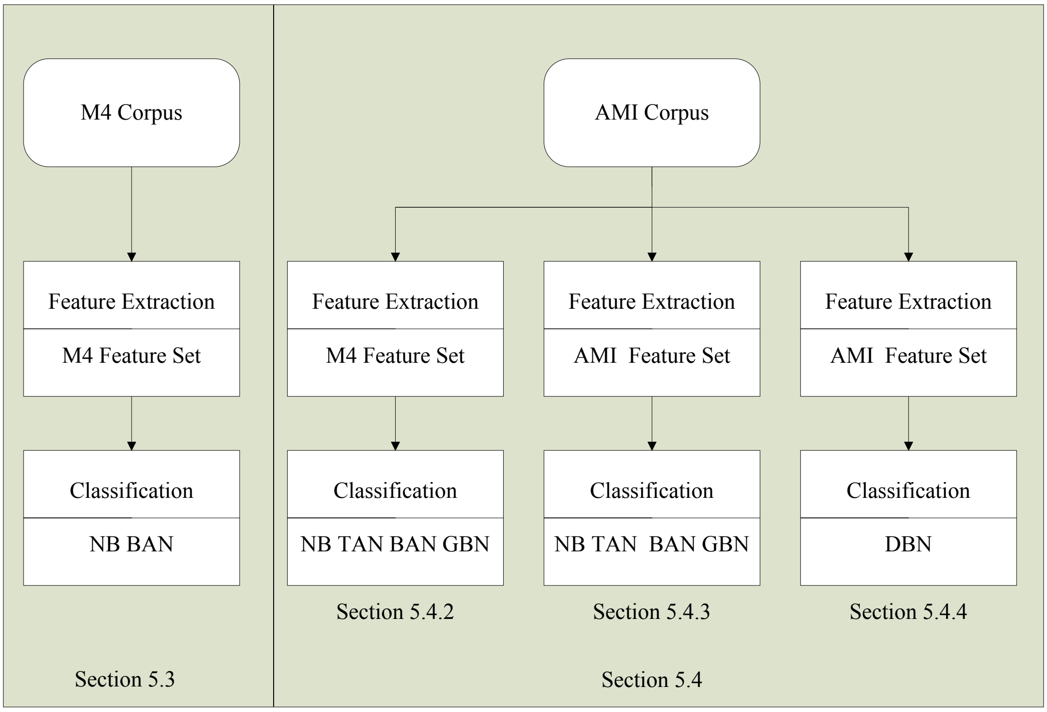


Figure 5.1: Experimental setup and an overview of the structure of this chapter

5.1 Tools

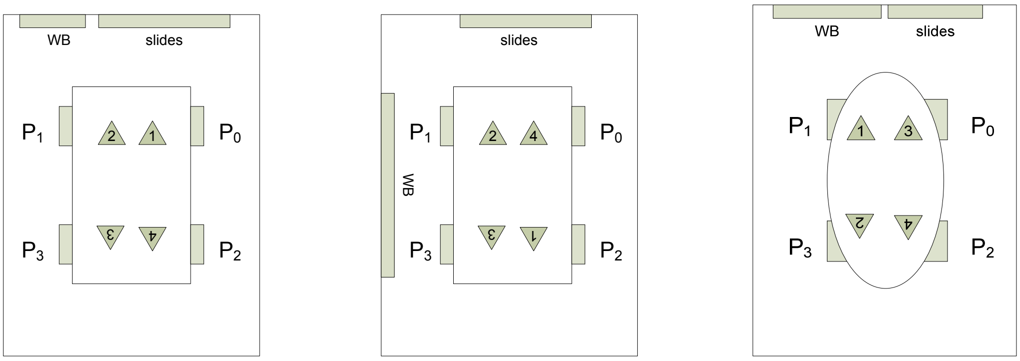
Experiments with the static BN classifiers presented in this chapter were conducted using various BN classifier learning algorithms implemented in WEKA (Witten and Frank, 2000) whereas experiments with the DBN classifiers were performed using the Bayes Net Toolbox (BNT) for MATLAB (Murphy, 2001). Additionally, we conducted experiments with the static BN classifiers using BNT in order to compare performances of dynamic and static BN classifiers.

As both meeting corpora are stored in the NXT stand-off XML format, we made use of the NXT Query Language to extract relevant features from various annotation layers (see Section 3.1). For the experiments on the M4 data, we created a number of data sets using the NXT Search engine. The queried results were stored in the Excel format and further processed using Visual Basic for Applications, an implementation of Visual Basic that is built into Microsoft Excel. For the experiments on the AMI data, we developed a NXT based application, Feature Extractor, implemented in Java that is employed for feature extraction and the creation of various data sets. Generated data sets are stored in the WEKA file format - Attribute-Relation File Format (ARFF)(Witten and Frank, 2000, p.49).

5.2 Classification task

In a dialogue situation, which is an event that lasts as long as the dialogue act performed by the speaker in that situation, the class variable is the addressee of the dialogue act (**ADD**). We define addressee classifier to identify for each *relevant dialogue act* whether the act is addressed to a particular individual or to a group. For addressee classification, irrelevant dialogue acts in the M4 schema are those dialogue acts that are marked as Unlabeled whereas in the AMI schema are those dialogue acts that are marked as Backchannel, Fragment, Stall or Other. These dialogue acts were not labeled with addressee information.

As discussed in Chapter 3, the IDAP, TNO and Edinburgh meeting rooms are relatively similar but differ in shape, construction and placement of the recording devices. As a result, meeting participants identified as A, B, C and D in the AMI corpus are located at different seating positions in different meetings rooms. Since the impact of gaze behavior on addressee prediction is affected by the seating arrangement of meeting participants, it is important to have a consistent identification of meeting participants across different meeting rooms. Therefore, in our model, participants in the meetings are identified based on the seating positions that are determined relative to the presentation screen in each meeting room: $\mathtt{P_x}$ denotes the participant at seating position $\mathtt{x}$, where $\mathtt{x} \in \{0, 1, 2, 3\}$ (see Figure 5.2). Clearly, this representation is plausible if participants do not switch seating positions during a meeting. In our data set, participants do not change seating positions. The mapping between participant's labels A, B, C and D that are related to speech channels in the AMI corpus and participant labels introduced in our model is performed using the resource file *meetings.xml*, which is available as a part of the AMI corpus. The file contains, among other things, the mapping between meeting participant labels, roles and audio/video channel IDs. The participant labels in the M4 corpus `Px`, which denotes the speaker at the channel `x`, coincide with labels introduced here for the purpose of addressee identification.



(a) IDIAP Room (b) Edinburgh Room (c) TNO Room

Figure 5.2: AMI Instrumented Meeting Rooms - the numbers in the triangles denote close-up camera IDs

Using the notation introduced in this section, the following class values are distinguished: individual `Px` where $\mathtt{x} \in \{0, 1, 2, 3\}$ and `Group`. As group and subgroup addressing are not distinguished in the AMI annotation schema, both types of addressing are marked as `Group`. However, the M4 annotation schema makes a distinction between addressing the whole group and addressing a subgroup of participants. Since there are only a few instances of subgroup addressing in the M4 data and annotators failed to agree on the subgroup categories, we excluded those instances from the data set. Therefore, the `Group` class value encompasses the whole audience in the M4 experiments whereas in the AMI experiments, it refers to any group of participants. Furthermore, both schemas allow for dialogue acts to be labelled with the Unknown/Unclassifiable addressee tag which denotes that annotators could not determine who is being addressed. These instances of dialogue acts were employed for deriving contextual information used for predicting the addressee of the dialogue act at hand, as will be explained later in this chapter.

5.3 Addressee classification on the M4 data

The goals of the experiments presented in this section are:

1. to find relevant features for addressee classification in meetings using information obtained from multi-modal resources - gaze, speech and conversational context

2. to explore to what extent the performances of classifiers can be improved by combining different types of features obtained from these resources

3. to investigate whether information about meeting context modeled in terms of meeting activities can aid the performances of addressee classifiers

4. to compare performances of the NB and BAN classifiers for the task of addressee prediction over various feature sets.

5.3.1 Feature set

To identify the addressee of a dialogue act we initially used three sorts of features: conversational context features (later referred to as contextual features), utterance features and gaze features. Additionally, we conducted experiments with an extended feature set including a feature that conveys information about meeting context.

Contextual features Contextual features provide information about the preceding *relevant* dialogue acts. We experimented with using information about the speaker, the addressee and the type of the immediately preceding dialogue act regardless of whether it is by the same or a different speaker (**SP-1**, **ADD-1**, **DA-1**) as well as information about the related dialogue act (**SP-R**, **ADD-R**, **DA-R**). A related dialogue act is the dialogue act that is the a-part of an adjacency pair with the current dialogue act as the b-part. The information about the speaker of the current dialogue act (**SP**) has also been included in the contextual feature set. If a dialogue act is unrelated **SP-R**, **DA-R** and **ADD-R** are assigned to NULL value. Similarly, if the previous relevant dialogue act does not exist corresponding contextual information is marked as NULL. Furthermore, the same value is assigned to **ADD-1** and **ADD-R** if the relevant dialogue act is marked with the Unknown addressee tag.

Utterance features The utterance feature set comprises the following set of binary lexical features:

- **PP** - does the utterance contain personal pronouns “we” or “you”, both of them, or neither of them?
- **PPA** - does the utterance contain possessive pronouns or possessive adjectives (“your/yours” or “our/ours”), their combination or neither of them?
- **IP** - does the utterance contain indefinite pronouns such as “somebody”, “someone”, “anybody”, “anyone”, “everybody” or “everyone”?
- **Name-P$_\mathtt{x}$** does the utterance contain the name of participant $\mathtt{P_x}$ where $\mathtt{x} \in \{0, 1, 2, 3\}$?

Utterance features also include information about the utterance’s conversational function (**DA-Type**) and information about utterance duration, that is, whether the utterance is short or long (**Short**). In our experiments, an utterance is considered as a short utterance, if its duration is less than or equal to 1 sec. The value set for the **DA-Type** feature and in line with that for **DA-1** and **DA-R** features includes the complete set of relevant DA tags that was used in the corpus creation (see Section 3.3.2). As discussed in Section 5.2, irrelevant dialogue acts are marked as Unlabelled.

Gaze features We experimented with a variety of gaze features. In the first experiment, for each participant $\mathtt{P_x}$ we defined a set of features in the form $\mathbf{P_x}$**-looks-**$\mathbf{P_y}$ and $\mathbf{P_x}$**-looks-NT** where $\mathtt{x}, \mathtt{y} \in \{0, 1, 2, 3\}$ and $\mathtt{x} \neq \mathtt{y}$; $\mathbf{P_x}$**-looks-NT** represents that participant $\mathtt{P_x}$ does not look at any of the participants. The value set represents the number of times that participant $\mathtt{P_x}$ looks at participant $\mathtt{P_y}$ or looks away during the time span of the utterance: `zero` for 0, `one` for 1, `two` for 2 and `more` for 3 or more times. In the second experiment, we defined a feature set that incorporates only information about the gaze direction of the current speaker (**SP-looks-**$\mathbf{P_x}$ and **SP-looks-NT**) with the same value set as in the first experiment.

Meeting context As to meeting context, we experimented with different values of the feature that represents the meeting actions (**MA-Type**). First, we used a full set of speech based meeting actions that was applied for the manual annotation of the meetings in the corpus: monologue, discussion, presentation, whiteboard, consensus and disagreement. As the results on modeling group actions in meetings presented in McCowan et al. (2003) indicate that consensus and disagreement were mostly misclassified as discussion, we have also conducted experiments with a set of four values for **MA-Type**, where consensus, disagreement and discussion meeting actions were grouped in one category marked as discussion.

5.3.2 Models, data set and evaluation method

Models Addressee classification on the M4 data has been performed by means of the NB and BAN classifiers. BAN has been chosen as a representative of more general BN classifiers in comparison to NB classifiers. Furthermore, we limited the maximal number of parents for each node in the BAN structure to 3. For learning the BAN structure, we applied the K2 algorithm explained in Chapter 4. The algorithm requires an ordering of the observable features; different ordering leads to different network structures. We conducted experiments with several orderings regarding feature types as well as with different orderings regarding features of the same type. In all orderings, the contextual features were first in the list and variations were based on the the positions of the remaining features in the set. The obtained classification results for different orderings were nearly identical. Parameters of the network were estimated using MAP estimates by setting all hyperparameters α_{ijk} to 0.5 (see Section 4.2.1, equation (4.21)).

Data set After we had discarded the instances labeled with `Unknown` or subgroup addressee tags, there were 781 instances left available for the experiments[1]. The distribution of the class values in the selected data is presented in Table 5.1.

[1] According to the M4 addressee annotation schema, dialogue acts marked as Unlabeled are assigned the Unknown addressee tag

Group	P_0	P_1	P_2	P_3
40.20%	13.83%	17.03%	15.88%	13.06%

Table 5.1: Distribution of class values - M4 data set

Evaluation method Performances of the addressee classifiers on the M4 data were evaluated in terms of accuracy. To evaluate performances of the addressee classifiers per class value we made use of precision, recall and F-measure. The performance measures are estimated using the stratified 10-fold cross validation (see Section 4.5).

5.3.3 Results and discussions

Experiments without meeting context

The performances of the classifiers are measured using different feature sets. First, we measured the performances of classifiers using utterance features, gaze features and contextual features separately. Then, we conducted experiments with all possible combinations of different types of features. As a simple baseline we used the classifier which always predicts the majority class. As shown in Table 5.1 the classifier tags all instances in the data set as addressed to the group which is correct in 40.20% of cases.

Table 5.2 summarizes the accuracies of the NB and BAN classifiers (with 95% confidence interval) for different feature sets (1) using gaze information of all meeting participants (Gaze ALL) and (2) using only information about speaker gaze direction (Gaze SP).

	NB		BAN	
Feature sets	**Gaze ALL**	**Gaze SP**	**Gaze ALL**	**Gaze SP**
All Features	78.10 (±2.90)	78.49 (±2.88)	81.05 (±2.75)	82.59 (±2.66)
Context	68.12 (±3.27)		73.11 (±3.11)	
Utterance+SP	52.50 (±3.50)		52.62 (±3.50)	
Gaze+SP	64.53 (±3.36)	59.02 (±3.45)	66.45 (±3.31)	62.36 (±3.40)
Gaze+SP+Short	65.94 (±3.32)	61.46 (±3.41)	67.73 (±3.28)	66.45 (±3.31)
Context+Utterance	72.21 (±3.14)		76.82 (±2.96)	
Context+Gaze	74.90 (±3.04)	77.59 (±2.92)	79.00 (±2.86)	80.03 (±2.80)
Utterance+Gaze+SP	69.78 (±3.22)	68.63 (±3.25)	70.68 (±3.19)	70.04 (±3.21)

Table 5.2: Accuracies of the BAN and NB classifiers using gaze information of all meeting participants (Gaze All) and using speaker gaze information (Gaze SP). The results for the feature sets without gaze features are summarized in the columns Gaze ALL

The results show that the BAN classifier outperforms the NB classifier for all feature sets, although the difference is significant only for the feature sets that include contextual features. Furthermore, both classifiers significantly outperform the baseline over all feature sets.

For the feature set that contains only information about gaze behavior combined with information about the speaker (Gaze+SP), both classifiers perform significantly better when exploiting gaze information of all meeting participants. In other words, when using solely focus of visual attention to identify the addressee of a dialogue act, listeners' focus of attention provides valuable information for addressee prediction. The same conclusion can be drawn when adding information about utterance duration to the gaze feature set (Gaze+SP+Short), although for the BAN classifier the difference is not significant. For all other feature sets, the classifiers do not show a significant difference in the accuracies when including or excluding the listeners gaze information. Moreover, both classifiers perform better using only speaker gaze information in all cases except when combined utterance and gaze features are exploited (Utterance+Gaze+SP).

The BAN and NB classifiers show the same changes in the performances over different feature sets. The results indicate that the selected utterance features are less informative for addressee prediction compared to contextual features (BAN:73.11%; NB:68.12%) or features of gaze behavior (BAN:66.45%, NB:64.53%). The results also show that adding the information about the utterance duration to the gaze features, slightly increases the accuracies of the classifiers (BAN:67.73%, NB:65.94%). This confirms findings presented in (Bakx et al., 2003) which have shown that some improvement in addressee detection in the context of mixed human-human and human-computer interaction can be achieved by combining utterance duration with gaze. Combining the information from the gaze and speech channels significantly improves the performances of the classifiers (BAN:70.68%; NB:69.78%) in comparison to performances obtained from each channel separately. Furthermore, higher accuracies are gained when adding contextual features to the utterance features (BAN:76.82%; NB:72.21%) and even more to the features of gaze behavior (BAN:80.03%, NB:77.59%). As is to be expected, the best performances are achieved by combining all three types of features (BAN:82.59%, NB:78.49%), although not significantly better compared to combined contextual and gaze features.

Likewise using information about the speaker of the a-part of an adjacency pair for detecting the addressee of its b-part, information about addressee of the b-part can be employed as a useful cue for identifying the speaker of the a-part and thus the related dialogue act. Therefore, we also explored how well the addressee can be predicted excluding information about the related utterance. The best performances are achieved by combining speaker gaze information with contextual and utterance features (BAN:79.39%; NB:76.06%). The decrease in the performance of about 3% can be considered as small since contextual information provided by the related dialogue act is shown to be a strong indicator for addressee detection: using only contextual information obtained from the related dialogue act both classifiers scores are around 66%. Therefore, it can be concluded that remaining contextual, utterance and gaze features capture most of the useful information provided by the related dialogue act.

Precision, recall and F-measures presented in Tables 5.3 and 5.4 indicate that each classifier performs almost equally well over all class values according to F-measures. The confusion matrixes of the best performed BAN and NB classifiers presented in Tables 5.5 and 5.6 show that most misclassifications were between addressing types (individual vs.

Class	Precision	Recall	F-measure
P_0	0.765	0.843	0.802
P_1	0.831	0.850	0.840
P_2	0.829	0.823	0.826
P_3	0.918	0.765	0.834
Group	0.821	0.831	0.826

Table 5.3: Evaluation per class value for the BAN classifier using utterance, contextual and speaker gaze features

Class	Precision	Recall	F-measure
P_0	0.781	0.759	0.770
P_1	0.789	0.789	0.789
P_2	0.858	0.734	0.791
P_3	0.831	0.676	0.746
Group	0.751	0.847	0.796

Table 5.4: Evaluation per class value for the NB classifier using utterance, contextual and speaker gaze features

group): each P_x was more confused with Group than with P_y where $x, y \in \{0, 1, 2, 3\}$. The same type of confusion is observed between human annotators regarding addressee annotation (see Section 3.3.4). Out of all misclassified cases for each classifier, individual types of addressing were, on average, misclassified with addressing the group in 73% cases for NB, and 68% cases for BAN.

	P_0	P_1	P_2	P_3	G
P_0	91	2	4	0	11
P_1	4	113	2	0	14
P_2	6	3	102	0	13
P_3	3	1	1	78	19
G	15	17	14	7	261

Table 5.5: Confusion matrix for the BAN classifier; rows - ground truth values; columns - classified values; G - Group

	P_0	P_1	P_2	P_3	G
P_0	82	5	6	0	15
P_1	4	105	1	2	21
P_2	5	2	91	1	25
P_3	1	4	1	69	27
G	13	17	7	11	266

Table 5.6: Confusion matrix for the NB classifier; rows - ground truth values; columns - classified values; G - Group

Similar to using Kappa for measuring agreement between human annotators on how they classify dialogue act segments into a set of addressee categories (see Section 3.2), Kappa can be used to assess how well an automatic annotator, that is, an automatic classifier, agrees with a human annotation of the material, which is taken as the ground truth. Kappa values for the best performing BAN and NB classifiers are 0.71 and 0.77, respectively. These values are in line with the Kappa values achieved by trained human annotators reported in Section 3.3.4 ($\kappa = 0.70$ and $\kappa = 0.81$).

Experiments with meeting context

We examined whether information about meeting context can aid the classifiers' performances. First, we conducted experiments using the six value set for the **MA-Type** feature. Then, we experimented with the reduced set containing four types of meeting actions (see Section 5.3.1). The accuracies of the classifiers obtained by combining the **MA-Type** feature with contextual, utterance and gaze features are presented in Table 5.7.

	NB		BAN	
Feature sets	**Gaze All**	**Gaze SP**	**Gaze All**	**Gaze SP**
MA-6+All	78.74	79.90	81.82	82.84
MA-4+All	78.23	79.13	81.69	83.74

Table 5.7: Accuracies of the NB and BAN classifiers using the MA-Type feature combined with the utterance, gaze and contextual features

The results indicate that adding meeting context information to the initial feature set improves the classifiers' performances slightly, but not significantly. The highest accuracy (83.74%) is achieved using the BAN classifier by combining the four-values **MA-Type** feature with contextual, utterance and the speaker's gaze features.

5.3.4 Summary of findings

In summary, we can draw the following conclusions from the experimental results presented in this section:

- Contextual information aids classifier performances over gaze and utterance information

- Selected utterance features are the most unreliable cues for addressee prediction.

- Listeners' gaze direction provides useful information only in the situation where gaze features are used alone.

- Combinations of features from various resources improve classifiers' performances in comparison to performances obtained from each resource separately

- The highest accuracies are achieved by combining contextual and utterance features with the speaker's gaze directional cues.

- Addressee classifiers show a small decrease in the accuracy when information about the related dialogue act is excluded from the contextual feature set, which indicates that the most useful information provided by the related dialogue act is captured by remaining contextual, utterance and gaze features.

- Addressee classifiers show little gain from information about meeting context

- The BAN classifier outperforms the NB classifier over all feature sets, although the difference is significant only when contextual features are exploited.

- Addressee classifiers mostly misclassified individual and group addressing

5.4 Addressee classification on the AMI data

The experiments on the AMI data were conducted with the following goals:

- to explore how well the addressee of a dialogue act can be predicted on the AMI data using the feature set defined in Section 5.3.1 (*M4 feature set*)
- to investigate whether the performances of addressee classifiers can be improved using a modified set of contextual, utterance and gaze features (*AMI feature set*)
- to explore the impact of meeting context modelled in terms of topical structure on the classifiers' performances.
- to examine the effect of using knoweldge about the roles participants perform in meetings on the performances of the addressee classifers
- to compare performances of the TAN and GBN classifiers in addition to the NB and BAN classifiers for the task of addressee prediction on the AMI data over various feature sets.
- to explore the impact of using the classified instead of the hand-annotated value for the addressee of the immediately preceding dialogue act on the classifiers' performances.

5.4.1 Models, data sets and evaluation methods

Models Addressee classification on the AMI data was performed by means of several static BN classifiers as well as by means of the DBN classifier. As to static classifiers, we experimented with the NB, TAN, BAN and GBN classifiers. The NB and BAN classifiers have also been employed for the comparison of classification results on the M4 and AMI data sets.

Structures of the static BN classifiers were learned from the data whereas structures of the DBN classifier were designed based on the learned static structures. For the comparison of the DBN and static BN classifiers, we experimented not only with the learned but also with the specified structure of the static BN classifiers. The K2 algorithm was applied for learning the structure of the BAN and GBN classifiers using the same ordering of feature types as in the M4 experiments. The structure of the TAN classifier was learned using the algorithm developed by Friedman et al. (1997)(see Section 4.3). Regarding the BAN and GBN classifier, we experimented with several thresholds for the maximal number of parents for each node. In addition to three as the parent threshold that was employed on the M4 data, we conducted experiments with setting the maximal number of parents to four and five. As the accuracies obtained for different parent thresholds were nearly identical, we report the classification results for the best performing BAN and GBN classifiers.

For learning parameters of the static BN classifiers in the experiments where structures were learned from the data, we used the same method as in the M4 experiments: the MAP

Data	Description	Total	P_0	P_1	P_2	P_3	Group
A set	- IS1006d	5380	13.61%	11.08%	9.28%	9.80%	56.25%
B set	+ IS1006d	6077	14.30%	10.70%	9.13%	11.14%	54.73%

Table 5.8: AMI data sets

estimation with $\alpha_{ijk} = 0.5$. For learning parameters of the DBN classifier, we applied the EM algorithm with uniform Dirichlet priors on network parameters (see Section 4.2.1). The MAP algorithm with uniform Dirichlet priors was employed for learning parameters of the static BN classifiers with the fixed structure.

Data sets Only a small part of the AMI scenario-driven collection has been annotated with addressee information. For the experiments presented in this chapter, we selected 14 meetings that were annotated with addressees and focus of attention[2]: ES2008a, TS3005a, IS1000a, IS1001a, IS1001b, IS1001c, IS1003b, IS1003d, IS1006b, IS1006d, IS1008a, IS1008b, IS1008c, IS1008d. Most of the selected meetings were recorded in the IDIAP meeting room. As the IS1006d meeting was not annotated with all types of information that we used in some of our experiments, we created two data sets to experiment with: the **A set** that excludes the IS1006d meeting and the **B set** that contains all 14 meetings. For each data set, the distribution of class values is given in Table 5.8. The total numbers of instances presented in the column Total denote the total numbers of relevant dialogue act segments that are marked with a class label.

Evaluation methods For evaluating performances of the static BN classifiers, we performed stratified 10-fold cross validation on the A set. For the experiments with DBNs, we made use of the larger B data set because the features selected for those experiments were annotated for all meetings included in the B set. Moreover, we divided the data set into 5 folds, 4 of which contained 3 meetings and one contained 2 meetings, ensuring that folds contain approximately similar number of instances. Due to uneven distribution of addressee values across the meetings, it was not feasible to specify the data set partition into n folds containing the meetings that completely satisfy the stratification criterion. In our partition, the average difference between the distribution of the addressee values in corresponding training and test folds is 2.5% with the maximal difference of about 10% for the group addressee value in one of the folds. A detailed specification of the data set partition is given in Appendix A. To compare performances of the DBN and static BN classifiers, the static classifiers were also evaluated using 5-fold cross validation on the defined folds.

In addition to the overall accuracy, the detailed accuracies per class value have been estimated in terms of precision, recall and F-measure. As discussed in Section 5.2, relevant dialogue acts marked with the Unclassifiable addressee tag were employed for deriving

[2] There are several meetings in the AMI corpus annotated with addressee information for which focus of attention annotations are not available

contextual information used for predicting the addressee of the dialogue act at hand. In the static case, those instances were removed from the data set after the contextual information obtained from them had been encoded in the feature set for the dialogue act that follows. In the dynamic case, however, the instances of relevant dialogue acts marked with Unclassifiable addressee labels were not removed from the data set as in that case the preceding dialogue act for the dialogue act that follows would be modeled differently than in the static case: the dialogue act that precedes the dialogue act marked as Unclassifiable would be marked as the immediately preceding dialogue act for the dialogue act that follows the Unclassifiable one. In the DBN model, addressee values for the Unclassifiable dialogue acts were treated as missing. However, instances of dialogue acts in the test set with missing addressee values were not considered in the estimation of the classifier's performance. In other words, the accuracies of the DBN classifier were calculated as the ratio between correctly classified test instances and the total number of the instances in the test set that were annotated with a class value.

5.4.2 Addressee classification using static BN classifiers - M4 feature set

The performances of the static BN classifiers on the AMI data were estimated using the initial feature set that was shown to score the highest accuracy on the M4 data. The set contains contextual, utterance and speaker gaze features described in Section 5.3.1.

As the AMI and M4 dialogue act tag sets contain different categories, the features exploited for the experiments on the AMI data differ from the original feature set in the value set for the **DA** feature and in line with that for the **DA-1** and **DA-R** features. Similar to the experiments on the M4 data, the complete set of relevant dialogue act tags was employed as the value set for the **DA** feature on the AMI data (see Section 3.4.2). As discussed in Section 5.2, irrelevant dialogue act types for addressee prediction on the AMI data are Backchannel, Stall, Fragment and Other. The mapping between the M4 and AMI dialogue act tag sets is presented in Section 3.5.

In addition to the M4 schema where a dialogue act can be marked either as unrelated or as related to a previous dialogue act produced by a different speaker, a dialogue act in the AMI schema can also be related to a previous dialogue act produced by the same speaker or to something that is not expressed verbally (see Section 3.5). The contextual feature set encompasses only those related dialogue acts that are produced by a different speaker. In all other cases, the dialogue act is considered as unrelated.

Table 5.9 summarizes the accuracies of the static BN classifiers estimated on the AMI data using the M4 feature set. The baseline classifier, which tags all instances as addressed to the group, has an accuracy of 56.25%. The results show that the accuracies for the augmented NB classifiers and the GBN classifier do not differ significantly although the BAN classifier performs the best. Furthermore, all three classifiers outperform the NB classifier. To evaluate whether the difference in the accuracies between the NB classifier on one hand and the BAN and GBN classifiers on the other hand is significant we performed

Feature sets	NB	TAN	BAN	GBN
M4 Feature Set	74.33	76.36	76.73	76.60

Table 5.9: Accuracies of static BN classifiers estimated on the AMI using the M4 feature set

a pairwise T-test based on ten times 10-fold cross validation with the significance level $\alpha = 0.05$. The results presented in Appendix A show that the TAN, BAN and GBN classifiers significantly outperform the NB classifier when the M4 feature set is employed for the addressee prediction on the AMI data.

The results presented in Table 5.9 indicate a significant decrease in the performances of the NB and BAN classifiers on the AMI data in comparison to their performances on the M4 data (see Table 5.2, All Features, Gaze SP). As shown in Tables 5.3 and 5.10 as well as in Tables 5.4 and 5.11 this decrease in performances is reflected in significantly worse performances regarding the prediction of individual addressee values. However, both classifiers perform similarly on both data sets regarding the group classification.

Class	Precision	Recall	F-measure
P_0	0.719	0.697	0.708
P_1	0.735	0.708	0.721
P_2	0.684	0.681	0.683
P_3	0.729	0.740	0.734
Group	0.805	0.815	0.810

Table 5.10: Evaluation per class value for the BAN classifier using the M4 feature set

Class	Precision	Recall	F-measure
P_0	0.686	0.582	0.630
P_1	0.773	0.629	0.694
P_2	0.689	0.591	0.636
P_3	0.760	0.632	0.690
Group	0.754	0.849	0.799

Table 5.11: Evaluation per class value for the NB classifier using the M4 feature set

To investigate the reasons for the decrease in the performances on the AMI data regarding individual addressee values, we estimated how well the addressee of a dialogue act can be predicted using each feature type separately. The mean F-measure was used as an indicator of the classifiers' performances regarding the recognition of individual addressee values. The obtained results have shown that utterance and gaze features are even less effective cues for addressee prediction on the AMI data than on the M4 data.

When utterance features are used alone in combination with speaker information, the NB and BAN classifiers score considerably lower F-measures regarding individual addressing on the AMI data (NB: mean F-measure=0.172, BAN: mean F-measure=0.183) in comparison to the M4 data (NB: mean F-measure=0.288, BAN: mean F-measure=0.296). Considerably worse performances on the AMI data regarding the individual addressee values when using solely utterance features may be, among other reasons, due to the difference in value sets of the DA feature: the M4 dialogue act tag set provides more valuable information for distinguishing group from individual addressing as it contains, for example, acknowledgement or response categories that are on the M4 data in many cases addressed to a previous speaker. As discussed in Chapter 3, some of utterance conversational functions that are captured with the M4 response categories (e.g. agreement, disagreement) are marked in the AMI schema with the assessment category. This category, on the other hand, is a very broad category in the AMI schema. It captures other aspects of an utterance's conversational function (e.g. expressing an opinion), which causes this category to be, in the substantial number of cases, addressed to the group.

Likewise using solely utterance features, the NB and BAN classifiers show substantially worse performances in predicting individual addressee values on the AMI data (NB: mean F-measure=0.445, BAN:mean F-measure=0.407) than on the M4 data (NB: mean F-measure=0.581, BAN: mean F-measure=0.591) when gaze features are used alone in combination with speaker information. One of the reasons for this drop of effectiveness of the gaze features on the AMI data regarding identification of single addressed dialogue acts can be influenced by the presence of additional attention distracters in the meeting room such as remote control prototypes or laptops. Furthermore, the impact of gaze behavior on the classifier performances can also be influenced by the manner in which focus of attention was annotated in the AMI corpus: a person A is considered as looking at B, only when A looks at B's face. As an example, looking at a participant who is writing something on the whiteboard is annotated as looking at the whiteboard. Therefore, when a participant in the audience asks the participant at the whiteboard a question, gaze information alone does not provide valuable information for detecting the addressee of that question.

5.4.3 Addressee classification using static BN classifiers - AMI feature set

Feature set

In the previous section, we have shown that utterance and speaker gaze features as specified in the M4 feature set, are less reliable cues for addressee identification on the AMI data. Aiming to improve classification results, we defined a modified feature set of utterance, gaze and contextual features focusing mostly on better exploitation of contextual information as it was shown that conversational context contributes to the largest extent to addressee prediction. Additionally, we conducted experiments with an extended feature set that includes a feature that conveys information regarding the meeting context in terms of topical organization. Finally, we explored the impact of the background knowledge about the roles participants play in meetings modeled in terms of role features on the performances of the addressee classifiers on the AMI data.

Contextual features Two notions of conversational context - namely, local and global - have been employed for modelling contextual information obtained from the dialogue acts that precede the current one. In both cases, only relevant preceding dialogue acts were taken into account.

The **local context** encompasses contextual information obtained from the relevant dialogue acts from the same or a different channel that most recently precede the current dialogue act. In other words, it comprises n-grams of the preceding dialogue acts. In addition to the immediately preceding dialogue act (1-gram) that was modelled in the M4 feature set, we also experimented with the extended context that includes two (2-gram) and three (3-gram) preceding dialogue acts. Contextual information obtained from the i-th preceding dialogue act encompasses information about the speaker (**SP-i**), the addressee (**ADD-i**) and the type (**DA-i**) of that dialogue act.

As to the **global context**, we distinguished contextual information obtained from a previous turn from the contextual information obtained from the turn in progress. A turn is defined as a sequence of successive dialogue acts $\mathtt{DA_i}$ for $\mathtt{i} = \mathtt{1}, \ldots, \mathtt{N}$, produced by the same speaker that satisfy one of the following conditions:

- $\mathtt{start(DA_{i+1})} - \mathtt{end(DA_i)} = \mathtt{0}$
- $\mathtt{0} < \mathtt{start(DA_{i+1})} - \mathtt{end(DA_i)} \leq \mathtt{T}$, where $\mathtt{T}$ is a defined threshold, and there are no "turn-relevant" dialogue acts produced by other speakers that occur within the gap between $\mathtt{DA_i}$ and $\mathtt{DA_{i+1}}$. In our experiments, irrelevant dialogue acts for the definition of turns are those dialogue acts marked as Backchannel, Stall or Fragment.

The i-th preceding turn ($\mathtt{T_i}$) of the dialogue act $\mathtt{DA_x}$ is defined as a turn that contains the first relevant dialogue act $\mathtt{DA_y}$ preceding $\mathtt{DA_X}$ that is not part of previous turns $\mathtt{T_1}, \ldots, \mathtt{T_{i-1}}$. Turns containing only irrelevant dialogue acts are considered as irrelevant turns.

The second condition in the definition of a turn specifies three types of silence defined in Chapter 2: pause, gap and lapse. If the condition is satisfied, the silence is considered as a pause and $\mathtt{DA_{i+1}}$ is included in the current turn. If the difference is greater than $\mathtt{T}$, the silence is classified as a lapse if there are no turn-relevant dialogue acts produced by different speakers occurring within the gap, otherwise it is classified as a gap. In both cases, $\mathtt{DA_{i+1}}$ is marked as the first dialog act of the next turn. It is to be noted that the definition of the turn also supports simultaneous speech.

As discussed in Chapter 2, pauses in ordinary conversations are brief. In meetings, a speaker, however, can make longer pauses while, for example, working on a laptop or while drawing something on the whiteboard. In this situation, according to Edelsky's (1981) distinction between floor and turns, the speaker is having the floor that consists of several turns. In our experiments, this type of having the floor is considered as one turn. Empirical analysis of the data shows that the maximal duration of silences of this type of holding the floor was around 5 sec; most of them are actually less than 3 sec. Therefore, we experimented with $\mathtt{T}$=5 sec.

Contextual information of a preceding turn encompasses information about the speaker, the addressee and the type of the relevant dialogue act of that turn which most recently preceded the current dialogue act. Contextual information of the current turn comprises information about the addressee and the type of a preceding relevant dialogue act of that turn. We conducted a number of experiments with various *window-sizes* regarding the number of preceding turns as well as regarding the number of preceding dialogue acts within the same turn. Additionally, we explored the performances of addressee classifiers when previous turns of the current speaker were both included and excluded from the contextual feature set. The results reported in this section are achieved using two contextual feature sets:

- **C11**- contains contextual information obtained from the immediately preceding turn (**SP-T-1**, **ADD-T-1**, **DA-T-1**) and contextual information obtained from immediately preceding dialogue act within the same turn (**ADD-1**, **DA-1**)
- **C21**-contains contextual information obtained from two preceding turns (**SP-T-1**, **ADD-T-1**, **DA-T-1**, **SP-T-2**, **ADD-T-2**, **DA-T-2**) and contextual information obtained from immediately preceding dialogue act within the same turn (**ADD-1**, **DA-1**)

In both cases, the preceding turns of the current speaker were taken into account. The reasons for choosing these two feature sets are two-fold: (1) addressee classifiers achieved the highest accuracies when those features are combined with gaze and utterance features and (2) the obtained results are comparable to those achieved using the selected n-gram local context features.

Information about the related dialogue act (**SP-R**, **ADD-R**, **DA-R**) and information about the speaker of the current dialogue act (**SP**) have also been included in the contextual feature set both when experimenting with the local context features and when experimenting with the global context features. Related dialogue acts are defined in the same way as in the M4 feature set (see Section 5.4.2).

For any contextual features included, the NULL value has been introduced to account for instances in which a previous dialogue act segment, as specified in the local or global context, does not exist. The same value is assigned to addressee contextual features that were marked with the Unclassifiable addressee tag.

Gaze features Based on the finding summarized in Section 5.3.4 that listeners' gaze direction provides valuable information for addressee prediction only in the situation when gaze features are used alone, we experimented solely with speaker gaze directional cues on the AMI data.

The experiments were conducted using two groups of gaze features. The first group consists of the features defined in the M4 feature set: **SP-looks-P**$_\mathtt{x}$ and **SP-looks-NT**, where $\mathtt{x} \in \{\mathtt{0},\mathtt{1},\mathtt{2},\mathtt{3}\}$; **SP-looks-NT** represents that the speaker does not look at any of the participants. The second group of features includes all categories that are labelled as gazed targets in the AMI schema: participants (**SP-looks-P**$_\mathtt{x}$), whiteboard (**SP-looks-WB**), presentation slides (**SP-looks-PS**) and table (**SP-looks-T**). As in the first feature group, **SP-looks-NT** is used to denote that the speaker does not look at any of the labelled gazed targets.

We also experimented with two different value sets for both groups of features. First, we defined gaze features as binary features that mark whether or not the speaker looks at the particular gazed target or whether or not he looks away during the time span of the current dialogue act. Then, we experimented with the value set that represents the number of times the speaker looks at a gazed target or looks away in the course of the current dialogue act. The extension of the target set to include other objects in the meeting room had an effect on the distribution of the speaker gaze over the targets. An analysis of the AMI data has shown that instances where the speaker looks three or more times at a particular gazed target occur less frequently in the data. Therefore, we defined the following value set: `zero` for 0, `one` for 1, `more` for 2 or more.

We have found that the addressee classifiers perform slightly better using the limited feature set. Furthermore, when gaze features are used alone in combination with speaker information higher accuracies have been achieved using the value set that denotes qualitative account of the number of times a feature occurs during the time span of the current dialogue act. However, when gaze features are combined with other types of features, the classifiers perform better using the binary gaze features. For that reason, binary features **SP-looks-P$_{\mathtt{x}}$** and **SP-looks-NT** have been employed for the experiments presented in this section.

Utterance features Using the available annotations of dialogue acts and named entities, we experimented with a variety of utterance features that are considered with the content, duration and the conversational function of the current dialogue act. Some of the features have already been introduced in the M4 feature set.

- **PP$ feature set** encompasses subjective and objective personal pronouns, possessive pronouns and possessive adjectives. It consists of the following binary features: **1.sing**, **1.pl**, **2.sing/pl** and **3.pl/sing**. For example, **1.pl** denotes whether or not the utterance contains “we”, “us”, “our” or “ours”. In the M4 feature set, **PP$** is partially defined with four-values **PP** and **PPA** features that contain information about `we` and `you` person categories.

- **IP**- whether or not the utterance contains indefinite pronouns such as “somebody”, “someone”, “anybody”, “anyone”, “everybody” or “everyone”?

- **ParticipantRef feature set** includes the features that mark reference to meeting participants:

  - **Name-P$_{\mathtt{x}}$**- whether or not the utterance contains the name of participant P$_{\mathtt{x}}$ where $\mathtt{x} \in \{\mathtt{0},\mathtt{1},\mathtt{2},\mathtt{3}\}$. In order to distinguish the usage of the name as an addressed term from other usages, we also included the **BeginOrEnd-P$_{\mathtt{x}}$** feature in the set. It denotes whether or not the name of participant P$_{\mathtt{x}}$ occurs at the beginning or at the end of the utterance.
  - **Role-P$_{\mathtt{x}}$** - whether or not the utterance contains the role of the participant P$_{\mathtt{x}}$.
  - **NameOrRole-P$_{\mathtt{x}}$** - whether or not the utterance contains the name or the role of participant P$_{\mathtt{x}}$. **NameOrRole-P$_{\mathtt{x}}$** is mutually exclusive with **Name-P$_{\mathtt{x}}$** as well as with **Role-P$_{\mathtt{x}}$**.

- **Short**- whether or not the utterance duration is less than or equal to 1 sec.

- **NumWords**- qualitative description of the number of words in the utterance: `one` for 1, `few` for 2, 3, 4 words, `many` for 5 or more words. As these **NumWords** and **Short** features provide almost redundant information, we decided to select one of those features in the final model.

- **Reflexivity** - whether or not the utterance is reflexive.

- **DA-Type** - the conversational function of the current dialogue act. In defining a value set for the **DA-Type** feature, we experimented with different groupings of the dialogue act categories. The results presented in this chapter were obtained using the following value set: `inform`, `assess`, `social`, `elicit`, `offer`, `suggest`, `comment-about-understanding`

Out of all listed utterance features, **PP$**, **DA-Type** and **NumWords** were shown to be the most informative when combined with selected contextual and utterance features. The results presented in this chapter are obtained using this subset of utterance features.

Meeting context Meeting context is modelled in terms of the **Topic** feature. Although the AMI topic segmentation schema allows topics to be nested up to several levels, we experimented only with top-level topics, which reflect largely the meeting structures based on the meeting scenario (see Section 3.4.2). Functional topics, as defined in the AMI schema, can also be labeled as top-level topics. As they reflect the actual process and flow of meetings (e.g opening, closing), they were also taken into account in modeling meeting context. Although the schema provides a pre-defined set of topic descriptions for top-level topics, annotators were allowed to introduce their own descriptions when necessary. However, we considered only pre-defined topic descriptions; all other descriptions were grouped into the `other` category.

The value set for the Topic feature contains the following descriptions: `agenda/equipment`, `opening`, `closing`, `project specification`, `new requirements`, $\mathtt{P_0}$`-present`, $\mathtt{P_1}$`-present`, $\mathtt{P_2}$`-present`, $\mathtt{P_3}$`-present`, `discussion`, `prototype presentation`, `prototype evaluation`, `project evaluation`, `costing`, `drawing` and `other`. Regarding the topics that refer to presentations, the AMI annotation schema contains the descriptions that refer to participant roles such as marketing expert presentation or industrial designer presentation. However, in the data processing step we mapped these values into corresponding values P_0-present, P_1-present, P_2-present, P_3-present incorporating in that way the background knowledge of the participant roles into the classification models.

Participant roles In addition to incorporating the background knowledge about the roles participants play in the AMI meetings in an implicit way by mapping some of the features or feature values defined in terms of participant roles into corresponding features or feature values specified in terms of the participants that play these roles (e.g. **Name-P**$_\mathbf{x}$ or **Topic**), we also modeled this knowledge in an explicit way by defining new features that bear information about participant roles.

The experiments were conducted using solely information about the speaker role modeled in two different ways. First, we introduced the **Dominant** feature which denotes whether or not the speaker is the participant with the dominant role in the meeting, that is, project manager. Second, we experimented with the **SP-Role** feature which marks one of four AMI scenario roles the speaker fulfils in a meeting: `PM`, `ID`, `UI` or `ME`. The motivation for using the Dominant feature is that the participant with the dominant role in a meeting is expected to address the whole audience on average more than is case with the other meeting participants. However, as discussed in Section 2.4, the leading role in the meeting can also be determined by the current meeting activity. For example, a presenter during the presentation can take over the leading role for that part of the meeting or a participant with a particular role may become the dominant speaker when a topic related to his work and knowledge is being discussed. For some types of activities defined in the AMI meetings, such as presentations, this type of information has already been encoded in the data processing step. Introducing the SP-Role feature, we aimed to investigate whether the information about the dominant role for other types of activities and topics can be extracted from the SP-Role feature. However, we have found that the knowledge about the particular role that the speaker performs in the meeting does not provide any additional information in comparison to the information provided by the Dominant feature. This can also be caused by the fact that the meeting context is modeled in terms of the Topic feature which bears information about meeting structure specified more in terms of meeting activities (e.g. discussion or opening) than in terms of the particular topic being discussed (e.g. look and usability, components and materials, trend watching). The latter is captured mostly with the subtopics specified in the AMI topic annotation schema. The results presented in this section were obtained using the Dominant feature.

Experiments with utterance, gaze and contextual features

Table 5.12 summarizes the accuracies of the static BN classifiers evaluated on the A set using utterance, gaze and local context features. The comparison of the accuracies of the BN classifiers for the 1-gram model and the accuracies obtained using the M4-feature set (see Table 5.9) shows that using the simplified set of utterance and gaze features as defined in the AMI feature set all BN classifiers score the accuracies similar to those obtained using the M4 feature set. Furthermore, the results show that when extending the local context, that is, when going from the 1-gram model to the 2-gram model, all classifiers show a slight improvement in performance, although the NB classifier gain the most (about 2%). However, further extension of the local context leads to a decrease in the accuracies for all classifiers except for NB. This may be due to increase in dimensionality of the models by adding more features, which requires more parameters to be estimated. For reliable estimates of the parameters, more data are needed.

To compare performances of the static BN classifiers for each feature set we performed a pairwise T-test based on ten times 10-fold cross validation with the significance level $\alpha = 0.05$. The results of the tests are presented in Appendix A. The results indicate that for all three feature sets there is no significant difference in the performances among

Context	NB	TAN	BAN	GBN
1-gram	73.96	76.91	77.10	76.36
2-gram	75.86	77.49	78.05	76.93
3-gram	76.56	77.36	76.99	76.91

Table 5.12: Accuracies of the static BN classifiers using utterance, gaze and local context features

the TAN, BAN and GBN classifiers. For 1-gram model, all three classifiers significantly outperform the NB classifiers. However, for the 2-gram model, the GBN classifier does not show a significant difference in the performance in comparison to the NB classifier. The further extension of the local context to the 3-gram model results in the classification accuracies of the TAN, BAN and GBN classifiers that do not differ significantly in respect to the classification accuracies of the NB classifier.

Tables 5.13 and 5.14 show the detail evaluation per class value and the confusion matrix for the best performing addressee classifier - namely, the BAN classifier which employs utterance, gaze and 2-gram context features.

Class	Precision	Recall	F-measure
P_0	0.713	0.754	0.733
P_1	0.727	0.767	0.746
P_2	0.687	0.687	0.687
P_3	0.728	0.782	0.754
Group	0.836	0.805	0.820

Table 5.13: Evaluation per class value for the BAN classifier using utterance, gaze and 2-gram context features

	P_0	P_1	P_2	P_3	G
P_0	552	7	9	7	157
P_1	9	457	11	10	109
P_2	9	17	343	14	116
P_3	5	7	8	412	95
G	199	141	128	123	2435

Table 5.14: Confusion matrix for the BAN classifier using utterance, gaze and 2-gram context features; rows - ground truth values; columns - classified values

The comparison of the performances of the BAN classifiers presented in Tables 5.13 and 5.10 indicates that the extension of the local context from the 1-gram model (Table 5.10, the M4 feature set) to the 2-gram model (Table 5.13) somewhat improves the classifier performances regarding individual addressee values. As presented in the confusion matrix, most misclassifications were between individual and group addressing. The same type of the confusion has been observed between human annotators on the AMI data (see Section 3.4.3). Furthermore, the best performing addressee classifier reaches $\kappa = 0.64$ whereas the worse performing addressee classifier - namely, the NB classifier which employs utterance, gaze and 1-gram context features (see Table 5.12) - scores $\kappa = 0.57$. These Kappa values are in line with pairwise Kappa values between human annotators who took part in the corpus creation $0.50 \leq \kappa \leq 0.63$ (see Section 3.4.3). The type of confusion as well as the Kappa values indicate that although agreement between human annotators on addressee annotation was somewhat low, automatic addressee classifiers did not show unexpected behavior with the respect to the type and amount of confusion when trained and evaluated on the hand annotated data.

The accuracies of the static BN classifiers obtained using the utterance, gaze and the global context featured set are summarized in Table 5.15.

Context	NB	TAN	BAN	GBN
C11	75.35	78.42	78.53	77.14
C21	76.13	78.81	78.81	76.67

Table 5.15: Accuracies of the static BN classifiers using the utterance, gaze and global context features

The classification accuracies obtained using the C11 context feature set combined with utterance and gaze features are comparable to the accuracies obtained using the 2-gram local context model. However, the C11 model contains the smaller number of features as the information obtained from the turn in progress exclude speaker information. On the other hand, for the first dialogue act of each turn, the C11 model provides the same information as the 1-gram model although represented with a reduced feature set. The same comparison holds for the C22 model, on the one hand, and 3-gram and 2-gram models, on the other hand.

The results presented in Table 5.15 indicate that modelling conversational context in a global way which distinguishes information obtained from the previous turns from the conversational information obtained from the turn in progress does not significantly change classifier performances in comparison to performances obtained using the local context (see Table 5.12). However, the TAN and BAN classifiers show the largest gain from modelling context in the global way: comparing the C21 model with the 3-gram model, the TAN and BAN classifiers increase their accuracies for 1.45% and 1.82%, respectively. The NB classifier, on the other hand, scores lower accuracies when the global context is employed for addressee detection than when the local context is considered. The results presented in Table 5.15 also indicate that extending the global context from the C11 model to C21 model increases slightly but not significantly the performances of the NB and augmented NB classifiers. This kind of the change in the performances regarding the augmented NB classifiers is different from the change in the performances that was observed when extending the local context from 2-gram model to 3-gram model (see Table 5.12).

Likewise comparing the performances of the static BN classifiers using the local context features, the pairwise T-test was performed for the comparison of the classifiers using the global feature set (see Appendix A). The results show that the augmented NB classifiers significantly outperform the NB and GBN classifiers for both feature sets. For the C11 model, the GBN classifier significantly outperforms the NB classifier. However, for for the C21 model, the difference in the classification accuracies between the NB and GBN classifiers is not significant.

We also explored how well the addressee classifiers perform on the AMI data when information about related dialogue act is excluded from the contextual feature set. The

Context	NB	TAN	BAN	GBN
Local context				
1-gram	71.52	74.57	75.50	74.09
2-gram	73.55	76.15	75.97	73.66
3-gram	74.13	75.97	75.39	72.49
Global context				
C11	72.66	76.43	76.04	74.96
C21	72.88	77.14	76.45	74.35

Table 5.16: Accuracies of the static BN classifiers using utterance features and gaze features combined with the contextual feature sets that exclude information about the related dialogue act

results of the experiments are given in Table 5.16. The results indicate that when contextual information obtained from the related dialogue act is not considered, all classifiers show a decrease in the performances for all feature sets. Comparing to classification results presented in Tables 5.12 and 5.15 for the corresponding feature sets, the decrease ranges from 1.3% for the TAN classifier using the 2-gram model to 4.4% for the GBN classifier using the 3-gram model with an average decrease of about 2%. As discussed in Section 5.3.2, the decrease of about 2% in the accuracies when information about related dialogue act is excluded can be considered as small indicating in that way that remaining contextual, utterance and gaze features cover the most useful information provided by the related dialogue act.

The TAN and BAN classifiers are the most resistant to excluding information about the related dialogue from the both local and global contextual feature sets. The GBN classifier, however, is the most affected when local context features are employed for addressee prediction. Furthermore, the drop in the accuracies of the GBN classifier when excluding information provided by the related dialogue act increases with the extension of the local context. This is somewhat surprising as it would be expected that the extension of the context could capture the valuable information provided by the related dialogue act since related dialogue acts in multi-party conversations are not always adjacent. However, modelling context in the global way provides more valuable information for addressee prediction in respect to the GBN classier when contextual features obtained from the related dialogue act are excluded in comparison to information provided by the local context without related dialogue act features. Likewise the GBN classifier, the augmented NB classifiers show better performances using the global representation of the context. The NB classifier, on the other hand, achieves considerably lower accuracies when using the global context features than when using the local context features. Analysis of the accuracies of the addressee classifiers per class value have shown that the decrease in the performances when information about related dialogue act is excluded is reflected in the significantly worse performances regarding the individual addressee value.

Disadvantages of the global context Modelling context in the global way has several disadvantages. First, the definition of the turn is dependent on the threshold parameter T which is employed for distinguishing three types of silence - gap, pause and lapse - i.e. for determining whether two successive dialogue acts produced by the same speaker with a silence in between are to be merged into one turn. Second, the determination of turns is affected by the reliability of manual or automatic annotation of turn-irrelevant dialogue acts. Third, for identifying the addressee of a speaker utterance in real-time applications, the sequential modeling approach which encompasses the n-gram contextual model is more appropriate. Therefore, the experiments presented in the following sections are based on the initial feature set that contains utterance, gaze and local context features.

Experiments with meeting context

Table 5.17 summarizes the accuracies of the static BN classifier obtained by combining the **Topic** feature with utterance, gaze and local context features.

Context	NB	TAN	BAN	GBN
1-gram	74.67	77.68	77.84	76.39
2-gram	76.73	77.94	77.64	77.01
3-gram	77.19	77.83	77.38	76.65

Table 5.17: Accuracies of the static BN classifiers using utterance, gaze, local context and Topic features

The results given in Table 5.17, compared to the results presented in Table 5.12, indicate that all addressee classifiers show a little gain from the information about meeting context modeled in terms of the Topic feature. The average gain is about 0.5% although the GBN classifier shows the smallest increase in the accuracies whereas the NB classifier gains the most from the topic information. Moreover, the BAN and GBN classifiers score slightly lower accuracies when the Topic feature is combined with the 2-gram and 3-gram models, respectively, than when the Topic feature is not taken into consideration. Similar impact of the meeting context on the classifier performances has been observed on the M4 data where the meeting context feature was modelled in terms of meeting activities. The results also indicate that when the meeting context information is employed for addressee prediction, the augmented NB classifiers are somewhat less sensitive to the extension of the conversational context when the meeting context is taken into account than in the case when it is not considered.

Experiments with participant roles

Table 5.18 shows the accuracies of the static BN classifiers achieved by combining the **Dominant** feature with the initial feature set containing utterance, gaze and local context features as well as with the extended feature set that encompasses the Topic feature. Moreover, we present here the results for the 2-gram context features as it was previously shown that most of the addressee classifiers score the highest accuracies using the 2-gram contextual feature set.

Feature set	NB	TAN	BAN	GBN
2-gram+Dominant	75.89	77.47	78.18	77.03
2-gram+Topic+Dominant	76.65	78.10	77.90	77.01

Table 5.18: Accuracies of the static BN classifiers using utterance, gaze, 2-gram context, topic and role features

Comparing the classification accuracies presented in Table 5.18 and the accuracies presented in Tables 5.12 and 5.17 for the 2-gram model, we can conclude that knowing whether the speaker plays the dominant role in a meeting or not, does not provide valuable information for the addressee prediction on the AMI data set in addition to information provided by utterance, gaze, contextual and topic features. This can be due to fact that about 88% of the dialogue acts in the data for which the speaker has the dominant role in the meeting are dialogue acts performed by P_0.

Another way to include the background knowledge about the participant roles in the classification models is to redefine the classification task in a way that addressee classifiers identify whether a participant with a particular role in the meeting or a group has been addressed by the speaker. Obviously, this approach does not make it possible to incorporate the knowledge about the seating arrangement in a classification model since participants who play the same role in different meetings are in a number of cases seated at different positions.

5.4.4 Addressee classification using DBN classifiers

In the experiments with the static BN classifiers, we investigated how well the addressee can be predicted using, among other features, information regarding the addressees of preceding dialogue acts. The following notions of a 'preceding dialogue act' were taken into account:

- the immediately preceding dialogue acts from the same or a different channel
- the last dialogue act of a previous turn that precede the current dialogue act
- the preceding dialogue act of the same turn
- the related dialogue act

Considering the addressee information of a preceding dialogue act as given, we aimed to estimate the upper bound for the performances of the BN classifiers for the task of addressee prediction over the M4 and AMI feature sets. In this section, we present results on addressee classification where the classified instead of hand annotated value for the addressee of a preceding dialogue act is used as a feature for identifying the addressee of the current dialogue act. For that purpose, the DBN classifier has been employed.

Seeing this as a sequential modelling task, we experimented with the complete contextual information provided from the immediately preceding relevant dialogue act from the same or different channel (**SP-1**, **ADD-1**, **DA-1**). As shown in the experiments with the static BN classifiers, the augmented BN classifiers and the GBN classifier score higher accuracies for the 2-gram model than for the 1-gram model although the improvement is not significant. In DBNs, the information is propagated through the wider history, that is, the information provided from the dialogue act segment at time slice T-2 influences the addressee of the dialogue act at slice T implicitly through the information provided by the dialogue act at slice T-1. Since there is no significant improvement between 1-gram and 2-gram models, explicitly modeling relations between dialogue act segments at T-2 and T may just unnecessarily increase the complexity of the network and thereby require more data for learning the parameters.

Besides excluding information about the addressee of the related dialogue act (**ADD-R**) from the contextual feature set as it represents the class variable, we also excluded the information about the type of the related dialogue act (**DA-R**), experimenting in this way only with the information about the speaker of the related dialogue act (**SP-R**). Additionally, we conducted experiments without the **SP-R** feature aiming to develop the model for addressee prediction which outcome can be used for the detection of the related utterance. Utterance and gaze features as defined in the AMI feature set have also been employed for addressee classification using DBN.

As discussed in Section 5.4.1, the structures of the DBN classifier were designed based on the learned static structures, which provide visual insight into relationships between features and the class variable as well as between features themselves. We experimented with a variety of static structures, and their modifications, that were learned using the feature set selected for the experiments presented in this section by performing 10-fold cross validation on the B-set. The highest accuracies, which are reported in this section, are obtained using the designed structure presented in Figure 5.3.

As shown in Figure 5.3, the static classifiers have learned dependencies between the speaker looking at participants seated at the same side of the table (e.g. **SP-looks-P$_0$** and **SP-looks-P$_2$**). In this way, the background knowledge regarding the seating arrangement is incorporated in the network structure. The structure also captures dependencies among contextual features as well as among utterance features, in particular **DA-1** and **DA** features, between two adjacent slices. We did not model dependencies between corresponding gaze features of two adjacent dialogue act segments due to two main reasons. First, as the gaze information is modeled in terms of the speaker gaze features, a speaker change has the effect that the dependencies between corresponding gaze features (**SP-looks-P$_x$(t-1)** and **S-looks-P$_x$(t)**) does not model the probability that the same target is gazed in two adjacent slices as is the case if the speaker change did not occur. Second, as the sequence of the *relevant* dialogue acts is modeled in DBN, there can be a time gap in between two adjacent relevant dialogue acts which may have an influence on the speaker gaze behavior. It is to be noted, that **SP-R** is modeled as a static variable whereas the **SP-1** feature is

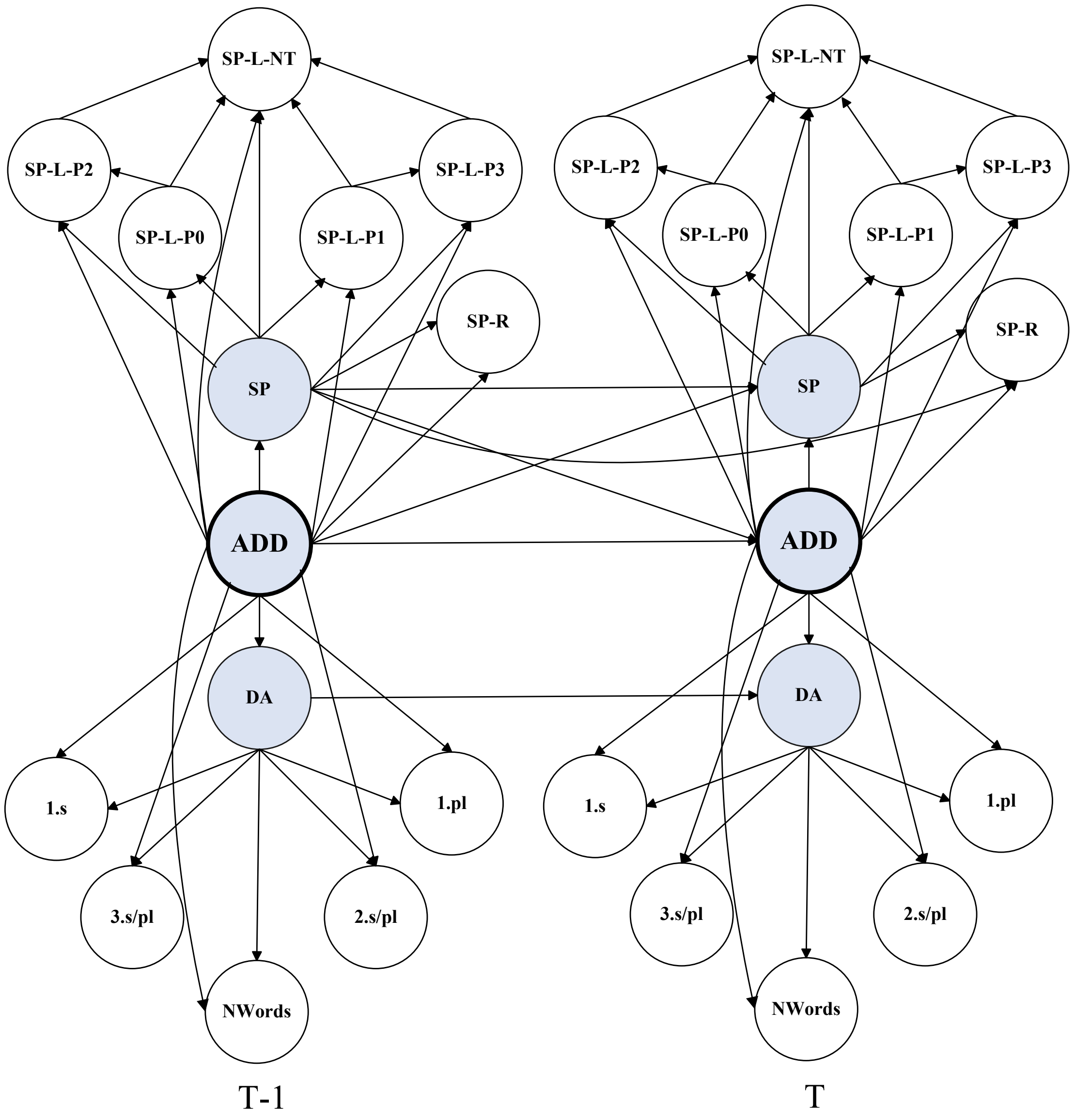


Figure 5.3: The structure of the DBN classifier

modeled as a dynamic variable in the network (**SP(t-1)**).

As discussed in Section 5.4.1, we considered contextual information provided by relevant dialogue acts that are marked with the Unclassifiable addressee label for the prediction of the addressee of the dialogue act at hand. In the dynamic case, as opposed to the static case, we did not exclude dialogue acts marked with the Unclassifiable addressee label from the data set as it would affect some other dialogue act to precede the current one and thus contextual information to be derived from that dialogue act. Therefore, a larger number of instances is employed for training the DBN classifier with some of them having missing addressee values. However, for evaluating the performances of the DBN classifier, we considered only those instances in a test set that were annotated with a class value (see Section 5.4.1). About 4% of all instances in the B set were labelled as Unclassifiable. Due to incompleteness of the data set, the EM algorithm with uniform Dirichlet prior has been employed for computing the MAP estimates of the network parameters (see Section 4.2.1).

The performances of the DBN classifier were evaluated by performing 5-fold cross validation on the B set (see Section 5.4.1). To gain an insight into how well the DBN classifier performs in comparison to the static BN classifiers, we evaluated the performances of the static BN classifiers using the same evaluation method. The static BN classifiers were developed in two different ways:

1. **static condition** - using the K2 algorithm for the structure learning and the MAP estimator with $\alpha_{ijk} = 0.5$ for parameters learning.

2. **dynamic condition** - using the the MAP algorithm with uniform Dirichlet priors for learning the parameters of the network with the fixed structure - the structure presented in Figure 5.3 transformed into a static network.

The static classifier with the fixed network structure similar to the one presented in Figure 5.3 was designed in a way that the addressee node is treated as a root node in the network. Furthermore, the addressee node was defined as a parent of all feature nodes. Since feature nodes form an arbitrary graph, the addressee classifier with such a structure is defined as the BAN classifier. Therefore, out of all static BN classifiers employed in Sections 5.4.3 and 5.4.2, we report only the results for the BAN classifier for the static condition. Table 5.19 summarizes classification accuracies of the DBN classifier as well as of the static BN classifiers for both static and dynamic conditions.

Feature set	**DBN**	**BAN(1)**	**BAN(2)**
All Features	71.83	75.63	75.58
All Features\{SP-R}	68.25	73.21	73.45

Table 5.19: Accuracies of the DBN classifier and the static BN classifiers under (1) static and (2) dynamic conditions

The results indicate that for both feature sets, the static BN classifiers significantly outperform the DBN classifier. Furthermore, both static and dynamic addressee classifiers significantly outperform the baseline classifier which always predicts the majority class (54.73%). Although the classification results for the dynamic and static BN classifiers are not quite comparable due to different treatment of the Unclassifiable addressee value, from the presented results we can conclude that the usage of classified instead of the hand annotated value has a negative impact on the classifier performances. Furthermore, both static and dynamic BN classifiers show a decrease in the performances when information about the speaker of the related dialogue act is excluded from the contextual feature set (DBN: about 3.5%, BAN: less than 2.5%). These results are comparable to the results obtained using the static BN classifiers on the A set when information about related dialogue act was excluded from the contextual feature set (see Section 5.4.3). It is to be noted, that the static BN classifier with the designed structure shows similar performances as the static BN classifier which structure was learned on each training fold.

Further analysis of the misclassified data instances has shown that the DBN classifier failed in a considerable number of cases to detect a change in addressing within a turn, especially when the speaker changes from talking to an individual to talking to a group and vice versa. This can be due to the fact that this type of change in many cases is not marked by the change in gaze behavior but by specific features of the utterance content that are not captured with the selected feature set.

5.4.5 Summary of findings

In summary, the following conclusions can be drawn from the experiments performed on the AMI data:

M4 and AMI:

- Using the M4 feature set containing utterance, contextual and speaker gaze features, the addressee classifiers score significantly lower accuracies on the AMI data than on the M4 data.
- The addressee classifiers show significantly worse performance regarding recognition of individually addressed dialogue acts on the AMI data, which is caused by insufficient information provided by utterance and gaze features.

Features:

- Extending the conversational local context from the 1-gram model to the 2-gram model improves slightly the performances of the static BN classifiers on the AMI data. However, further extension of the conversational context decreases the performances of all static BN network classifiers except of the NB classifier which show a little gain from the extension of the conversational context.
- Modeling conversational context in the global way does not significantly change classifier performances in comparison to performances obtained using the local context.

However, the augmented NB classifiers show the largest gain from modeling context in the global way.

- Addressee classifiers, both static and dynamic, show a small decrease in the performances when contextual information obtained from the related dialogue act is excluded from the contextual feature set. This indicates that remaining contextual, utterance and gaze features cover the most useful information for addressee prediction provided by the related dialogue act.
- Addressee classifiers show a little gain from information about meeting context modeled in terms of the Topic feature
- Information about the speaker role has no significant impact on addressee prediction when combined with utterance, gaze and contextual features as well as with the Topic feature.

Models:

- For all classifiers, the accuracies are significantly higher compared to the baseline.
- Augmented NB classifiers show the best performances over all feature sets.
- The NB classifier performs significantly worse than the augmented NB and GBN classifiers for the 1-gram model and significantly worse than the augmented NB classifiers for the 2-gram model. In all other cases, there is no significant difference in the accuracies among addressee classifiers when local context features are employed.
- The augmented NB classifiers significantly outperform the NB and GBN classifiers for both the C11 and C22 models. For the C11 model, the GBN classifier significantly outperforms the NB classifier. For the C21 model, the NB and GBN classifiers do not show significant difference in the performances.
- Static BN classifiers which use as a feature the hand annotated value for the addressee of the preceding dialogue act significantly outperform the DBN classifier which employs the classified value for the addressee of the preceding dialogue act.
- Addressee classifiers on the AMI data show the same type of misclassifications as human annotators: individual versus group.

5.5 Can addressee classification be further automated?

Experiments presented in this chapter were conducted using a set of manually annotated utterance, gaze, conversational context and meeting context features. Some of the features can be easily extracted (e.g **NumWords**) whereas another require complex computational modelling for their automatic detection (e.g. **DA-Type**, **SP-R**, **SP-Looks-P$_x$**, etc.). Moreover, we assumed that the dialogue act segment boundaries are given considering in that way the addressee classification task as a task of assigning the correct addressee labels to pre-segmented units. However, the ultimate goal for the application of addressee classification models in real systems is to perform this process fully automatically starting from audio and video signals. It requires, on one hand, that dialogue act segment boundaries as well as the utterance features and some of the contextual features are extracted from the output of an automatic speech recognizer (ASR) and on the other hand, that gaze features are estimated from visual information. Furthermore, most of the features can be better detected by combining multimodal - audio and video - cues.

Several issues that arise at this point are: (1) whether there are technologies available for automatic recognition of the features for addressee identification in the context of meetings (2) if so, what is the quality of the extracted features and (3) to what extent the quality of automatic feature extraction decreases the performances of the addressee classifiers achieved using the hand-annotated features.

In this section, we give answers to the first two questions by reviewing some existing work done on automatic recognition of the features used for the experiments with DBN (see Section 5.4.4) in the context of meetings. Furthermore, we highlight the existing work performed on the AMI data. However, the third question is not addressed in the scope of this thesis and it remains an open issue for the future work on automatic addressee modeling in face-to-face meetings.

5.5.1 An overview of the existing work on automatic detection of the features used in the DBN addressee classification model

Automatic identification of contextual features

Regarding contextual features, the main issue is how to automatically identify contextual information regarding the related dialogue act. To our knowledge, there is not much work done on automatic adjacency pairs identification in multi-party dialogues. Recently, Galley et al. (2004) reported the results on AP detection in multi-party meeting dialogues using the maximum entropy ranking model. The results indicate that given the b-part of an adjacency pair, the speaker of the a-part can be detected with accuracy of 90.12% using a set structural, durational and lexical features. This accuracy is achieved using "backward-looking" and "forward-looking" features. However, when excluding forward looking features, which concern the closest utterance of the potential speaker of the a-part that follows the b-part, the model performs worse (86.99%). It is to be noted, that the classification task for the experiments presented in (Galley et al., 2004) concerns the detection of the speaker of the a-part without identifying to which dialogue act of that speaker the b-part is related: the basic unit of analysis is a spurt which represents a period of speech that has no pauses greater than 0.5 sec. Therefore, for the experiments with the DBN presented in Section 5.4.4 we excluded the DA-R feature from the contextual feature set in addition to the ADD-R feature that was primarily excluded for the purpose of sequential modelling.

Automatic identification of utterance features

Recent development of the ICSI and AMI meeting corpora that are annotated with dialogue act (DA) information, has enabled significant work to be done on automatic dialogue act recognition in meeting dialogues. The DA recognition tasks consists of two subtasks - segmentation and classification - that can be performed either sequentially or jointly. In the sequential approach, segmentation and classification are implemented as two independent modules with the classification module using as its input the output of the segmentation module. The joint approach, however, implements both tasks simultaneously. Most of the past work on DA recognition was focused on DA classification using the hand segmented dialogues as the input. Recently new approaches have been proposed for fully automatic DA recognition encompassing both sequential and joint tasks. As a dialogue act segment is the basic unit for addressee classification and a dialogue act category is used as a feature for addressee prediction, the output of a DA recognition system should be used as an input for the addressee classification model.

So far, almost all reported work on DA recognition was performed on the ICSI meeting corpus (Janin et al., 2004). However, there are several ongoing studies on automatic DA annotation of the AMI corpus that are being carried out by the several sites in the AMI project. Some preliminary results reported in the AMI deliverable[3] are presented here.

The ICSI meetings have been automatically annotated using the following DA tag sets:

- the original MRDA tag set (Dhillon et al., 2004)
- different variants of the MRDA "classmap 1" that contain five categories: statement, question, backchannel, floor, and disruption
- the MALTUS (Multidimensional Abstract Layered Tagset for Utterances) tag set (Clark and Popescu-Belis, 2004).

In contrast to the AMI dialogue act annotation schema, the MRDA and MALTUS schemas allow for an utterance to be marked with a label made up of one or more tags from the set. This, however, effects a large number of the possible labels to be used for the automatic DA tagging of meeting dialogues. As the MALTUS tag set represents an abstraction of MRDA which is made by grouping some of the MRDA tags into classes and assigning mutually exclusive constraints among them, the number of possible labels is significantly reduced in comparison to the MRDA schema: the number of possible DA labels in MRDA reaches several millions whereas in MALTUS it amounts to 770 (Popescu-Belis, 2003; Clark and Popescu-Belis, 2004). The decrease in the number of the possible labels may have an effect on the performances of a dialogue act classifier. The MRDA "classmap 1", on the other hand, is a one-dimensional tag set containing 5 DA tags: each

[3] AMI Deliverable - D5.2 Implementation and Evaluation Results

utterance is labeled with exactly one tag from the set. In that regard, it is comparable to the AMI tag set which contains 15 DA classes.

Tables 5.20 and 5.21 list the performances of some of the existing DA segmentation, DA classification and DA recognition (segmentation and classification) systems evaluated on the ICSI and AMI data. Giving an overview of the performances of the existing systems for the DA recognition task and its subtasks, we do not aim to compare those systems in terms of the models and features used. Our goal is to gain some impression of how difficult these tasks actually are since the performances of addressee classification models can be influenced by the imperfection of these systems, especially by DA segmentation systems.

Condition	System	DSER
Ref	1	40.8
	2	36.8
	3	66.5
ASR	1	49.4
	2	47.4

1: Ang et al. (2005) - ICSI
2: Zimmermann et al. (2006b) - ICSI
3: AMI Deliverable D5.2 - AMI

Table 5.20: An overview of the performances of DA segmentation systems in multi-party meetings

.

The performances of the DA classification systems reported in Table 5.21 are evaluated in terms of the classification error rate which is defined as the percentage of incorrectly classified DA segments. DA segmentation and DA recognition are usually evaluated using a range of metrics which refer to different units: words, boundaries and dialogue act segments. For an overview of the existing metrics we refer to (Ang et al., 2005; Zimmermann et al., 2006a). The performances for DA segmentation and DA recognition systems presented here are measured using two dialogue act based metrics that are introduced in (Zimmermann et al., 2006a): DSER (DA Segmentation Error Rate) and DER (DA Error Rate). The DSER metric measures the percentage of the misclassified DA segments, that is, the segments for which at least one of the boundaries does not correspond exactly to the reference boundaries. DER is a dialogue act based metric which not only requires for the candidate DA segment to exactly match the boundaries of the reference segment but also to be tagged with the correct DA label. Where available, the results are reported for both reference orthographic transcription (Ref) and the output of automatic speech recognizer (ASR).

The presented results show that automatic DA classification, DA segmentation and especially DA recognition are very difficult tasks to perform automatically. The results also indicate that using the ASR output instead of the true sequence of the spoken words has a negative impact on the performances of all three types of system. The comparison of the results for Classmap 1 and AMI tag sets show that the performances of the DA

Condition	System	MRDA	MALTUS	Class 1	AMI
Manual segmentation					
Ref	1		27.7	22.1	
	2	49.0	39.2	21.3	
	3			18.1	
	5				34.1
ASR	3			25.5	
Automatic segmentation					
Ref	3			54.4	
	4			51.0	
	5			61.4	83.4
ASR	3			64.3	
	4			62.6	
	5			72.1	

1: Clark and Popescu-Belis (2004) 2: Lesch et al. (2005)
3: Ang et al. (2005) 4: Zimmermann et al. (2006b)
5: AMI deliverable D5.2

Table 5.21: An overview of the performances of the DA classification and the DA recognition systems in meeting dialogues. The DA classification systems are evaluated in terms of classification error (Manual segmentation). DA recognition systems are evaluated using the DER metric (Automatic segmentation).

classification and DA recognition systems on the AMI data are lower than the performances on the ICSI data. This can be due to the increase in the number of DA tags that was used for the automatic annotation of the AMI data. However, as previously mentioned, the results reported on the AMI data are preliminary. The ongoing work is focused on the improvement of both DA segmentation as well as DA classification systems and in line with that DA recognition systems on the AMI data.

From all these, we may conclude that at this moment the technology is not quite ready for handling a fully automatic dialogue act recognition, and in line with that addressee classification system.

Automatic identification of gaze features

Since it is very difficult to record eye gazing of meeting participants, the information about visual focus of attention can be automatically induced from head orientation. Experimental results presented in (Stiefelhagen and Zhu, 2002) indicate that estimation of focus of attention based solely on head orientation achieve the accuracy of 88.7% in four-participants meetings. In their previous work, Stiefelhagen et al. (2002) presented a system for estimating focus of attention based on multimodal cues: gaze directions and sound resources. First, participants' gaze direction was estimated from their head orientation. The gaze detected estimations are then used to predict focus of attention given a head pose. The scored accuracy using this approach is 74%. Adding audio information to video information increased the accuracy to 76%.

Preliminary results on recognition of focus of attention based on the head pose orientation on the AMI data are reported in (Al-Hames et al., 2006). In contrast to the work reported in (Stiefelhagen and Zhu, 2002) that was restricted to recognition of meeting participants as focus of attention targets, the recognition task presented in (Al-Hames et al., 2006) was considered with the recognition of the extended focus of attention label set that includes also table, slide screen and unfocused label. The obtained classification rate was 68% and 47%. As the authors claim, the lower recognition results are mainly due to the usage of more complex setting and the extended label set. The current research on the focus of attention recognition on the AMI meetings is concerned with adaptation of the approach to include information from other modalities in order to improve the classification results.

5.6 Conclusions

In this chapter, we presented results on addressee classification in four-participants face-to-face meetings using several types of static BN classifiers as well as using the DBN classifier. The classifiers were evaluated on the M4 and AMI meeting corpora using features obtained from multiple resources: speech, gaze, conversational context, meeting context and background knowledge about participant roles. As the features used in the classification models are based on hand annotated information, the experiments presented in this chapter concern establishing the upper bounds for the task of addressee prediction in face-to-face meetings.

The M4 and AMI meetings have different properties ranging from the short, informal, discussion meetings scripted in terms of meeting activities to the longer, more formal, task oriented meetings structured according to the AMI scenario that elicit more natural behavior of meeting participants. These differences have an impact on addressing behavior of meeting participants as well as on the distribution of addressing types in meetings. The former is also influenced by the layout of a meeting room. In the AMI scenario, the rooms are equipped with additional attention distracters such as remote control prototypes or laptops. There are two parameters that we kept invariant for both corpora: seating arrangement and the number of participants. Conducting experiments on the M4 and AMI meeting corpora, we also explored how well the findings regarding the addressee classification in a less realistic meeting scenario can be generalized to a more natural meeting scenario.

The experiments on the M4 data were conducted with the goal to explore relevant features for addressee detection in face-to-face meetings and to investigate how well the NB and BAN classifiers perform for this type of task. We have found that contextual information aids classifiers' performances over utterance and gaze information. Furthermore, utterance features were shown to be the most unreliable cues for addressee prediction. The exploration of the impact of listeners' gaze information on the performances of the addressee classifiers', leads us to the conclusion that listeners' gaze direction provides useful information only in the situation where gaze features are used alone. The addressee classifiers reach the highest accuracies when combining utterance and contextual features with the speaker's gaze directional cues. Combining information about meeting context modeled in terms of the current meeting activity with the utterance, contextual and speaker gaze features improves slightly but not significantly the classifiers' performances. Furthermore, we have shown that the BAN classifier outperforms the NB classifier although the difference is significant only when contextual features are employed. Each classifier shows similar performances regarding the classification of individual and group addressee values.

To gain an insight into how well the addressee classifiers perform on the AMI data in comparison to the M4 data, we evaluated the performances of the NB and BAN classifiers on the AMI data using the M4 feature set. In addition to the NB and BAN classifiers, we evaluated performances of the TAN and GBN classifiers for the task of addressee prediction on the AMI data. The results of the experiments indicate a considerable decrease in the performances of the NB and BAN classifiers on the AMI data which is reflected in significantly worse performances regarding the identification of individually addressed dialogue acts. Further explorations of the reasons for this decrease in the accuracies have shown that utterance and gaze features are even less effective cues for distinguish individual from group addressing on the AMI data than on the M4 data.

In contrast to Vertegaal (1998) and Otsuka et al. (2005) findings, where it is shown that gaze can be a good predictor for addressee in four-participants face-to-face *conversations*, the results of our experiments on both data sets, indicate that in four-participants face-to-face *meetings*, gaze is less effective as an addressee indicator. This can be due to several reasons. First, they used different seating arrangements which is implicated in the organization of gaze. Second, our meeting environments contain attention distracters. Finally, during a meeting, in contrast to an ordinary conversation, participants perform various meeting activities which may have an effect on gaze as an aspect of addressing behavior.

Since conversational context provides the most useful information for addressee prediction, we explored whether the performances of addressee classifiers on the AMI data can be improved by better exploitation of the contextual information. The conversational context has been modeled in two ways: local and global. The local context concerns n-grams of the preceding dialogue acts from the same or different channel. The global context, on the other hand, distinguishes contextual information obtained from the preceding turns from the contextual information obtained from the turn in progress. From the results concerning the local context, an important conclusion to be drawn was that the extension of the local context to include not only the immediately preceding dialogue act but also the dialogue act that precede that one, slightly improves performances of all static BN classifiers although the NB classifier gains the most. However, further extension of the conversational context decreases the performances of all addressee classifiers with the exception of the NB classifier which shows a small increase in the accuracies. From the experiments concerning the global context, we found that modeling context in the global way does not significantly change the performances of the addressee classifiers in comparison to the performances obtained using the local context although the augmented NB classifiers gain the most from the global representation of the context.

Exploring the impact of meeting context on classification results, we found that the addressee classifiers show little gain from the information about meeting context modeled in terms of the Topic feature which largely reflects the meetings structure according to the AMI scenario. This finding is in line with the findings related to the impact of meeting context on the classifiers' performances on the M4 data. It was also shown that information about the speaker role is an irrelevant cue for addressee prediction on the AMI data when combined with contextual, gaze, utterance as well as with topic features.

The overall conclusion that can be drawn regarding the performances of the static BN classifier on the AMI data over various feature sets is that the augmented NB classifiers outperform the NB and GBN classifiers for the task of addressee prediction. Furthermore, addressee classifiers evaluated both on the M4 and on the AMI meeting corpora mostly had problems in distinguishing between individual and group addressing.

This chapter also addressed the issue of the further automation of the addressee classification process. The DBN classifier was employed for modelling the first step in this process which comprises the usage of the classified instead of the hand annotated value for the addressee of the preceding dialogue act. The comparison of the performances between the DBN and static BN classifiers has shown that using the classified instead of the hand annotated value has a negative impact on the classifiers' performances: the observed decrease in the performance is about 4%. Further automation of the classification process comprises (1) automatic segmentation of meeting conversations into dialogue act segments and (2) automatic extraction of the remaining contextual, utterance and gaze features for each of those segments. The overview of the existing work in corresponding research areas has shown that both processes require complex computational modeling which currently produce considerable error rates. In this thesis, however, we did not explore the impact of these errors on the performances of addressee classifiers.

As addressee information can be used as a useful cue for the detection of the related utterance, we estimated the performances of addressee classifiers when information about related dialogue act was excluded from the contextual features set. It was found that the static BN classifiers evaluated on both meeting corpora and the DBN classifier evaluated on the AMI data show a similar decrease in the accuracies (about 3%) when information about related dialogue act is not taken into account. Since this information is a strong indicator for addressee prediction, this decrease in the performances indicates that remaining contextual, utterance and gaze features cover the most useful information provided by the related dialogue act.

5.6.1 Future recommendations

The addressee classification models presented in this chapter are concerned with real-time addressee detection. Therefore, the experiments presented in this chapter were conducted using the "backward looking" features which result from the analysis of the conversational context preceding the current dialogue act. However, for off-line addressee identification that can be applied for automatic meeting processing applications such as meeting browsers, contextual information provided from the dialogue acts following the current one may provide useful information for addressee detection: in many situations, the addressed participants is the one who speaks next. As discussed in Chapter 2, the first pair part of an adjacency pair when addressed to a particular individual is used as a technique for the selection of that individual as the next speaker. Therefore, the future work on addressee identification for the application in off-line systems can be concerned with the exploration of the usefulness of the "forward looking" features for addressee prediction.

The future work can also be focused on the exploration of additional sources of information that may be valuable for addressee detection on the AMI data such as information about locations of meeting participants as well as about the activities they are performing. Examples include sitting, writing on the whiteboard, standing by the projector screen while giving a presentation or taking notes using a laptop. This information is partially encoded in the AMI individual action annotation schema. Unfortunately, at the time when the experiments presented in this thesis were being performed this information was not available to us. As discussed in Section 5.4.3, the limitation of the gaze for addressee identification on the AMI data was partially influenced by the lack of this knowledge. Therefore, it is advisable for the future work on addressee identification on the AMI data to investigate the impact of using information about participants' location combined with the extending feature set that includes not only participants as gazed target but also the remaining focus of attention targets labeled in the AMI schema - namely, whiteboard, slide screen and table.

As the addressee classifiers had problems in detecting the changes in addressing within a turn, it would be valuable for the further research on this task to explore whether there are specific and observable patterns in addressing behavior of a speaker as well as in the behavior of the other participants that mark these changes. In the case that such patterns are detected, identification and modeling of the features of the detected behavior may help in better recognition of the changes in addressing within a turn.

The addressee classification task defined in this thesis supports the flat classification models which treat each class value separately. However, the classification task can be decomposed into two classification problems: first, the problem of distinguishing individual from group addressing and then, in the case of individual addressing, the problem of distinguishing among meeting participants who is the one being addressed by the speaker. Taking into account that the addressee classifiers mostly had problems in distinguishing individual from group addressing, defining the addressee classification task in the hierarchical manner may help in a more accurate distinction between these two categories. Moreover, some features that were shown to be less useful discriminators in the non-hierarchical representation (e.g. utterance features) may be potentially better discriminators in the hierachical representation. Therefore, the future work may also focus on the hierachical addressee classification and its comparision with the flat classification process.

Chapter 6

Conclusions

This chapter summarizes the thesis and brings some recommendations for future research.

6.1 Research summary and conclusions

As specified in Chapter 1, one of the goals of the present study was to gain knowledge in how addressing is accomplished in face-to-face meetings, that is, to understand the phenomenon under study. Chapter 2 has been devoted to this research goal. To achieve this, we drew heavily on the findings of studies on the organization of multi-party interaction conducted mostly in the area of conversational and interaction analysis. We pursued our analysis from the viewpoint of a two-dimensional organization of conversation: organization of conversation as sequences of actions and organization of conversation across participants. From our point of view, a conversation proceeds as a sequence of acts participants perform in speaking, each of them having assigned its *participation framework*. Participation framework, as defined by Goffman (1981a), is an ensemble of all of the participation roles of all participants in a conversation at a particular moment of speech, one of the roles being addressee.

We have distinguished two forms of addressing: explicit and tacit addressing. The latter refers to a manner of addressing which recipients draw upon diverse features of the *content* and *context* of the act performed by the speaker. Speakers may use various addressing devices *to make clear to whom* they are addressing their speech such as address terms, gaze, deictic hand and head gestures. We have shown that gaze and gestures when used alone can be troublesome and weak addressing devices as they may serve various purposes in conversation other than addressing. Speakers also employ various devices *to make clear that someone is explicitly addressed* without revealing who that person is (e.g. the usage of "you"). These partial devices need to be enhanced with other sources of information to accomplish addressing.

In many occasions in face-to-face conversation, addressing is, as our study reveals, carried out not explicitly but tacitly. Two-dimensional organization of conversation -

sequential organization and organization across participants - is what provides the context for accomplishing addressing tacitly. In Chapter 2, we discussed two practices that sequentially organize conversational interaction: turn-taking design and adjacency pairs organization. It was shown that they are highly interrelated with addressing practices. On the one hand, features of turn-taking and adjacency pairs organization provide context for tacit addressing. On the other hand, addressing and the initiative part of adjacency pairs are used in combination in the selection of the next speaker.

Meetings as a special form of talk in interaction differ from conversations in many aspects. We have shown that structural organization of meetings and activities in which participants are involved during a meeting have an influence on addressing behavior. Moreover, in meetings many utterances are broadcast. In many situations, the content of what has been said is relevant to all participants present. This brings difficulties in distinguishing whether a speaker addresses his speech to a particular individual or to the group as whole especially in situations where no explicating addressing devices are employed by the speaker.

As presented in Chapter 3, human observers having all the information available - namely, audio, video and speech transcriptions of meeting conversations - had mostly difficulties distinguishing between individual and group addressing. Moreover, determining whether a particular individual or a group is being addressed was found to be a more difficult task for human observers of the AMI meetings than of the M4 meetings. This was partially influenced by the meeting types. The AMI meetings, on the one hand, are task-oriented meetings where participants function as a team concerned with making decisions regarding the design of a new TV remote prototype. In these meetings, many of the utterances are intended and relevant for all participants present. The M4 meetings, on the other hand, are scripted and less interactive meetings in which people are gathered together to discuss personal experiences regarding various issues such as plans for the next holiday, a book or a movie. An initial exploration of annotation of subgroup addressing on the M4 data, brings us to the conclusion that this type of addressing is difficult to distinguish from addressing the whole group. More qualitative research thus needs to be conducted in that respect. Therefore, in our computation approaches to addressee identification, we have focused only on determining whether a particular individual or a group is addressed.

The second goal of our study, as stated in Chapter 1, was to develop a computational model for automatic addressee identification in meetings. We have chosen a Bayesian Network approach to this problem. The experiments were conducted with two main goals: finding relevant features for addressee identification obtained from various resources and finding appropriate BN models for this type of task. The detailed summaries of findings and general conclusions regarding the outcomes of the experiments are given in Chapter 5 (Sections 5.3.4, 5.4.5, 5.6). The most prominent findings are listed here.

The most important conclusion regarding the features employed is that conversational context features are essential for addressee prediction in face-to-face meetings. Utterance features, which contain a set of lexical features extracted from the utterance content as well as the type of dialogue act performed with the utterance, have shown to be the least reliable cues for addressee detection. Compared to empirical findings presented in the literature, gaze features were found to be less reliable cues for addressee predication in face-to-face meetings than in ordinary conversations. Relying only on gaze information, addressee classifiers show significantly lower performances compared to performances obtained using contextual features. The combination of features from all three resources leads to the best performances of addressee classifiers. Somewhat surprising is the finding that information about meeting context does not provide useful information for addressee identification in addition to contextual, utterance and gaze features.

Various approaches were taken to the extension of the conversational context. Experiments conducted in that regard have shown different changes in the performances for different BN classifiers. In general, all BN classifiers gain in accuracy from the extension of the conversational context that encompasses recent conversational history (e.g. contextual information obtained from the immediately preceding turn). However, with further extension of the conversational history addressee classifiers show a slight improvement or no improvement at all.

We also explored the relevance of the present work in another pragmatic research focusing on multi-party dialogues - namely, identification of adjacency pairs. Contextual features we considered include, among other things, contextual information provided by the related utterance. Likewise using information provided by the related utterance for addressee detection, addressee information can be employed as a cue for identifying the related utterance. A small decrease in the accuracies of the addressee classifiers when information about related utterance is not taken into account lead us to the conclusion that the present work may be useful for adjacency pairs identification.

Our exploratory study has shown that BNs are effective computational models for addressee identification in meetings. Evaluating several static BNs classifiers described in Chapter 4, we found that over all feature sets augmented NB classifiers - namely, the TAN and BAN classifiers - achieve the best results. Furthermore, the TAN and BAN classifiers hardly differ in their performances. The highest accuracies of 83.74% and 78.81% for the BAN classifier reported on the M4 and the AMI data, respectively, indicate quite high upper bounds for this type of task taking into account the complexity and ambiguity of the phenomena modeled. Another important conclusion is that the BN annotators show similar performances to human annotators in terms of the misclassifications made: most confusions were between individual and group addressing. Moreover, agreement between the BN annotators and human annotators is in line with agreement achieved between human annotators on addressee classification.

The features used to model addressing are in a sense ideal because they are obtained from human-annotated data. As the first step in the automation of addressee detection process we proposed a DBN model based on the previously listed findings regarding both relevance of the features and appropriateness of the static BN models. The model uses a classified instead of a hand-annotated value of the preceding addressee for the identification of the current addressee. Moreover, the model was designed to include only those relevant features for which there are existing technologies to be recognized automatically. The results of the evaluation of DBN classifier on the AMI data have shown that it achieves a significantly lower accuracy (about 4%) than the static BN classifiers which make use of hand-annotated features only.

6.2 Future work

Given that making a distinction between individual and group addressing was most problematic for both human and BN annotators, more theoretical research needs to be conducted with respect to better understanding and clarification of these two types of addressing. This would result in providing more reliable annotation instructions for human coders and in line with that in better performances of addressee classifiers.

As already discussed in Section 5.6.1, the future work regarding the experiments presented in this thesis may include the exploration of additional sources of information such as location of meeting participants and non-verbal activities in which they are involved (e.g. writing on the whiteboard or taking-notes using a laptop). It would be worthwhile to investigate whether this information provides valuable resources for the better exploitation of gaze features for addressee prediction. In this thesis, we were concerned with on-line addressee detection based on the features which result from the analysis of the conversational context preceding the current dialogue act. Human annotators also relied on the discussion preceding the dialogue act being labeled. However, for off-line meeting processing, the use of forward-looking features may be beneficial because the addressed person is in many situations the one who speaks next.

Two parameters were kept invariant in our study: the number of participants and seating arrangement. As they influence addressing behavior of meeting participants, it is advisable to investigate the effects these parameters have on addressee identification. Since the M4 and AMI meeting corpora do not suffice for this research purpose, annotation of new meetings containing different numbers of participants and various seating arrangements is required.

The natural next step in this research should be concerned with automation of the addressee classification process by using automatically identified instead of manually annotated features. Because automatic addressee classification is based, in the first place, on the outcome of a dialogue act segmentation system, new evaluation metrics for the performances of addressee classifiers should be defined similar to those for the evaluation of automatic dialogue act classification (see Ang et al., 2005; Zimmermann et al., 2006a).

The addressee classifiers presented in this thesis were evaluated by minimizing misclassification errors without taking into account the cost of making wrong decisions. In some meeting oriented applications such as automatic meeting minutes, it may be of greater importance to predict individual addressee values more accurately than group addressee values. Furthermore, the cost of failing to recognize that someone is addressed may be greater than recognizing that someone is addressed when he actually was not. Therefore, cost-sensitive evaluation of addressee classifiers may be desirable for their applications in real systems.

Recent advances in the context of automatic meeting analysis have enabled the shift of research focus from conventional meetings to live meetings with remote participants and from off-line to on-line meeting analysis. The recently started AMIDA (Augmented Multiparty Interaction with Distance Access) project[1] represents a challenging shift in that direction. The quality of technology used in computer mediated communication certainly has an influence on which aspects of face-to-face conversation can be reproduced in computer mediated communication. The usage of inexpensive devices such as web cams with the limited resolution affects the size of the window in which a remote participant is visible. This has an effect on the accomplishment of those aspects of conversational interaction that are maintained through non-verbal communication channels such as turn-taking, grounding and addressing. With respect to addressing in the remote meeting scenario, several interesting issues arise: (1) which addressing mechanisms remote participants use to address participants in meetings and vice versa?, (2) do addressing patterns in the live meetings change when the remote participant actively participates in conversation?

[1] AMIDA: http://www.amiproject.org/amida-scientific-portal

Appendix A

Experimental data

B set partition

B data set contains all AMI meetings that were annotated with addressee and FOA information. It is described in Section 5.4. Table A.1 outlines the partition of the B data set into 5 folds used for the experiments with the DBN addressee classifier presented in Section 5.4.4

		Fold 1		**Fold 2**		**Fold 3**		**Fold 4**		**Fold 5**	
Test	P0	159	12.44%	181	15.08%	225	18.70%	129	10.69%	175	14.72%
	P1	116	9.08%	129	10.57%	108	8.98%	145	12.01%	152	12.78%
	P2	132	10.33%	93	7.75%	117	9.73%	106	8.78%	107	9.00%
	P3	169	13.22%	121	10.08%	188	15.63%	117	9.69%	82	6.90%
	Group	702	54.93%	676	56.33%	565	46.97%	710	58.82%	673	56.60%
	Total	1278		1200		1203		1207		1189	
Training	P0	710	14.79%	688	14.11%	644	13.21%	740	15.20%	694	14.20%
	P1	534	11.13%	521	10.68%	542	11.12%	505	10.37%	498	10.19%
	P2	423	8.81%	462	9.47%	438	8.99%	449	9.22%	448	9.17%
	P3	508	10.59%	556	11.40%	489	10.03%	560	11.50%	595	12.17%
	Group	2624	54.68%	2650	54.34%	2761	56.65%	2616	53.72%	2653	54.28%
	Total	4799		4877		4874		4870		4888	
Meetings		IS1003d IS1008b		IS1006b IS1008a IS1008d		IS1006d IS1008c IS1001a		ES2008a IS1000a IS1001b		IS1001c IS1003b TS3005a	

Table A.1: The partition of the B set into 5 folds used for 5-fold cross validation of addressee classifiers

Comparison between addressee classifiers on the AMI data

Table A.2 summarizes the results of the pairwise two-tailed t-tests that were performed to test the statistical significance of differences in performance between the NB, TAN, BAN, GBN classifiers over various feature sets. The results are obtained by performing 10 times 10-fold cross validations on the AMI data. The chosen significance level is $\alpha = 0.05$. The confidence limit for Student's t-distribution with 9 degrees of freedom for $\alpha = 0.025$ is 2.262.

Feature sets	**TAN-NB**	**BAN-NB**	**GBN-NB**	**BAN-TAN**	**GBN-TAN**	**GBN-BAN**
M4 features	4.03*	4.60*	3.84*	1.52	0.45	-0.97
1-gram	5.25*	5.51*	4.27*	0.43	-1.21	-1.71
2-gram	3.74*	3.29*	1.98	0.28	-1.59	-2.07
3-gram	1.86	0.83	0.45	-0.86	-1.30	-0.53
C11	6.46*	5.90*	3.26*	-0.62	-3.42*	-3.25*
C21	5.37*	4.82*	0.51	-0.34	-4.69*	-5.04*

Table A.2: Results of the pairwise two-tailed T-tests for each pair of BN classifiers based on 10 times 10-fold cross validations. The results that are significant at the significance level 0.05 are marked with (*)

Bibliography

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transaction on Automatic Control*, AC-19(6):716–723.

Al-Hames, M., Hain, T., Cernocky, J., Schreiber, S., Poel, M., Muller, R., Marcel, S., van Leeuwen, D., Odobez, J., Ba, S., Bourlard, H., Cardinaux, F., Gatica-Perez, D., Janin, A., Motlicek, P., Reiter, S., Renals, S., van Rest, J., Rienks, R., Rigoll, G., Smith, K., Thean, A., and Zemcik, P. (2006). Audio-visual processing in meetings: Seven questions and current AMI answers. In *Proceedings of Workshop on Machine Learning for Multimodal Interaction (MLMI)*, Washington, DC, USA.

Ang, J., Liu, Y., and Shriberg, E. (2005). Automatic dialog act segmentation and classification in multiparty meetings. In *Proceedings of 30th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Philadelphia, PA, USA.

Argyle, M. (1973). *Social Interaction.* London: Tavistock Publications.

Austin, J. L. (1962). *How to Do Things with Words.* Cambridge: Harvard University Press.

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

Bales, R. F. (1950). *Interaction Process Analysis: A method for the studyof small groups.* Cambridge: Addison-Wesley.

Banerjee, S., Rose, C., and Rudnicky, A. I. (2005). The necessity of a meeting recording and playback system, and the benefit of topic-level annotations to meeting browsing. In *Proceedings of the 10th International Conference on Human Computer Interaction*, Rome, Italy.

Banerjee, S. and Rudnicky, A. I. (2006). You are what you say: Using meeting participants' speech to detect their roles and expertise. In *Proceedings of the NAACL-HLT 2006 workshop on Analyzing Conversations in Text and Speech*, New York, NY, USA.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning.* Springer.

Bloomfield, L. J. (1946). *Language.* New York: Henry Holt and Co.

Bouckaert, R. (1995). *Bayesian Belief Networks: from Construction to Inference.* PhD thesis, University of Utrecht, The Netherlands.

Boyen, X. and Koller, D. (1998). Tractable inference for complex stochastic processes. In *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence*, Madison, WI, USA.

Bunt, H. C. (2000). Dynamic interpretation and dialogue theory. In Taylor, M., Neel, F., and Bouwhuis, D., editors, *The Structure of Multimodal Dialogue, Volume 2*, pages 139–166. Amsterdam: John Benjamins.

Burger, S. and Sloane, Z. (2004). The ISL meeting corpus: Categorical features of communicative group interactions. In *Proceedings of 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., and Garrod, S. (2002). Seeing eye to eye: an account of grounding and understanding in work groups. *Bulletin of the Japanese cognitive science*, 9(1):1–20.

Carletta, J., Ashby, S., Bourban, S., M. Flynn, M. G., Hain, T., Kadlec, J., Karaiskos, V., W. Kraaij, M. K., Lathoud, G., Lincoln, M., Lisowska, A., McCowan, I., Post, W., Reidsma, D., and Wellner, P. (2005a). The AMI meeting corpus: A pre-announcement. In Renals, S. and Bengio, S., editors, *Proceedings of Workshop on Multimodal Interaction and Related Machine Learning Algorithms (MLMI)*, Edinburgh, UK. Springer-Verlag LNCS.

Carletta, J., Evert, S., Heid, U., and Kilgour, J. (2005b). The NITE XML Toolkit: Data Model and Query Language. *Language Resources and Evaluation Journal*, 39(4):313–334.

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

Carletta, J., Kilgour, J., O'Donnell, T., Evert, S., and Voormann, H. (2003). The NITE Object Model library for handling structured linguistic annotation on multimodal data sets. In *Proceedings of the EACL Workshop on Language Technology and the Semantic Web (3rd Workshop on NLP and XML, NLPXML-2003)*, Budapest, Hungary.

Cheng, J., Bell, D., and Liu, W. (1998). Learning Bayesian Networks from data: An efficient approach based on information theory. Technical report, University of Alberta.

Cheng, J. and Greiner, R. (1999). Comparing Bayesian network classifiers. In *Proceedings of the 15th Annual Conference on Uncertainty in Artificial Intelligence (UAI-99)*, pages 101–108, San Francisco, CA, USA.

Chickering, D. M. (1996). Learning Bayesian networks is NP-complete. In Fisher, D. and Lenz, H., editors, *Learning from Data: Artificial Intelligence and Statistics V*, pages 121–130. Springer-Verlag.

Chickering, D. M., D.Heckerman, and Meek, C. (2004). Large-sample learning of Bayesian networks is NP-hard. *Journal of Machine Learning Research*, 5:1287–1330.

Choudbury, T., Rehg, J. M., Pavlovic, V., and Pentland, A. (2002). Boosting and structure learning in Dynamic Bayesian Networks for audio-visual speaker detection. In *Proceeding of 16th International Conference on Pattern Recognition (ICPR)*, Quebec, Canada.

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

Clark, H. H. (1996). *Using language*. Cambridge: Cambridge University Press.

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

Clark, H. H. and Schaefer, F. E. (1992). Dealing with overhearers. In *Arenas of language use*. Chicago: University of Chicago Press.

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

Cooper, G. F. (1990a). Bayesian belief-network inference using recursive decomposition. Technical Report KSL 90-05, Knowledge Systems Laboratory, Medical Computer Science, Stanford University.

Cooper, G. F. (1990b). The computational complexity of probabilistic inference using Bayesian belief networks. *Artificial Intelligence*, 42(3):393–405.

Cooper, G. F. and Herskovits, E. (1992). A bayesian method for the induction of probabilistic networks from data. *Machine Learning*, pages 309–347.

Cremers, A. H. M., Hilhorst, B., and Vermeeren, A. P. O. S. (2005). "What was discussed by whom, how, when and where?" personalized browsing of annotated multimedia meeting recordings. In *Proceedings of HCI International*, Nevada, USA.

Dagum, P. and Luby, M. (1993). Approximating probabilistic inference in Bayesian belief networks is NP-hard. *Artificial Intelligence*, 60(1):141–153.

Dechter, R. (1996). Bucket elimination: A unifying framework for probabilistic inference. In *Proceedings of 12th Conference on Uncertainty in Artificial Intelligence*, Portland, OR, USA.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society (Series B)*, 39(1):1–38.

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

Duda, R. O. and Hart, P. E. (1973). *Pattern classification and scene analysis.* New York: John Wiley and Sons.

Duncan, S. J. (1972). Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, 23(2):283–292.

Duncan, S. J. and Niederehe, G. (1974). On signalling that it's your turn to speak. *Journal of Experimental Social Psychology*, 23:234–247.

Duranti, A. (1986). The audience as co-author: An introduction. *Text special issue*, 6(3):239–247.

Eco, U. (1986). Eine palettte von grautoenen. *die Zeit.*

Edelsky, C. (1981). Who's got the floor? *Language in Society*, 10:383–421.

Evert, S. and Voormann, H. (2002). The nite query language - version 2.1. Techical document. Available at `http://www.ltg.ed.ac.uk/NITE/documents/NiteQL.v2.1.pdf`.

Fleiss, J. (1981). *Statistical Methods for Rates and Proportions.* New York: John Wiley & Sons.

Friedman, J. H. (1997a). On bias, variance, 0/1 loss, and the curse of dimensionality. *Data Mining and Knowledge Discovery*, 1(1):55–77.

Friedman, N. (1997b). Learning belief networks in the presence of missing values and hidden variables. In *ICML '97: Proceedings of the Fourteenth International Conference on Machine Learning*, San Francisco, CA, USA.

Friedman, N., Geiger, D., and Goldszmidt, M. (1997). Bayesian network classifiers. *Machine Learning*, (29):131–163.

Friedman, N., Murphy, K., and Russel, S. (1998). Learning the structure of dynamic probabilistic networks. In *Proceedings of the Fourteenth Conference on Uncertainty in Artifical Intelligence (UAI)*, San Francisco, CA.

Fries, C. C. (1952). *The Structure of English.* New York: Harcourt, Brace and Company.

Galley, M., McKeown, K., Hirschberg, J., and Shriberg, E. (2004). Identifying agreement and disagreement in conversational speech: Use of Bayesian networks to model pragmatic dependencies. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, Barcelona, Spain.

Gatica-Perez, D., Zhang, D., and Bengio, S. (2005). Extracting information from multimedia meeting collections. In *Proceedings of 7th ACM SIGMM International Workshop on Multimedia Information Retrieval (MIR)*, Hilton, Singapore.

Gibson, D. R. (2003). Participation shifts: Order and differentiation in group conversation. *Social forces*, 81(4):1335–1381.

Goffman, E. (1963). *Behavior in Public Places: Notes on the Social Organization of Gatherings.* New York: Free Press of Glencoe.

Goffman, E. (1967). On Face-Work. In *Interaction Ritual: Essays on Face to Face Behavior*, pages 5–45. New York: Doubleday Anchor.

Goffman, E. (1981a). Footing. In *Forms of Talk*, pages 124–159. Philadelphia: University of Pennsylvania Press.

Goffman, E. (1981b). *Forms of Talk.* Philadelphia: University of Pennsylvania Press.

Goffman, E. (1981c). Radio talk. In *Forms of Talk*, pages 197–327. Philadelphia: University of Pennsylvania Press.

Goffman, E. (1981d). Replies and responses. In *Forms of Talk*, pages 5–77. Philadelphia: University of Pennsylvania Press.

Goffman, E. (1981e). Response cries. In *Forms of Talk*, pages 78–1123. Philadelphia: University of Pennsylvania Press.

Goffman, E. (1983). The interaction order (American Sociological Association 1982 Presidential Address). *American Sociological Review*, 48(1):1–17.

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

Goodwin, C. (1986). Audience diversity, participation and interpretation. *Text*, 6(3):283–316.

Goodwin, C. and Goodwin, M. H. (1990). Interstitial argument. In Grimshaw, A. D., editor, *Conflict Talk: Sociolinguistic Investigations of Arguments in Conversations*, pages 85–117. Cambridge: Cambridge University Press.

Greiner, R., Grove, A. J., and Schuurmans, D. (1997). Learning Bayesian nets that perform well. In *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence*, Providence, RI, USA.

Grossman, D. and Domingos, P. (2004). Learning Bayesian network classifiers by maximizing conditional likelihood. In *Proceedings of the 21st International Conference on Machine Learning (ICML'04)*, Banff, Alberta, Canada.

Haris, Z. (1951). *Methods In Structural Linguistics.* Chicago: Chicago University Press.

Haviland, J. B. (1986). 'Con buenos chiles': Talk, targets and teasing in Zinacantan. *Text*, 6(3):249–282.

Heckerman, D. (1995). A tutorial on learning with Bayesian networks. Techical report MSR-TR-95-06, Microsoft Research.

Heckerman, D., Geiger, D., and Chickering, D. (1995). Learning bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20:197–243.

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

Hollander, E. P. (1985). Leadership and power. In Lindzey, G. and Aronson, E., editors, *The handbook of social psychology.* New York: Random House, 3th edition.

Hymes, D. H. (1974). *Foundations in Sociolinguistics: An Ethnographic Approach.* Philadelphia: University of Pennsylvania Press.

Jaffe, J. and Feldstein, S. (1970). *Rhythms of dialogue.* New York: Academic Press.

Jaimes, A., Omura, K., Nagamine, T., and Hirata, K. (2004). Memory cues for meeting video retrieval. In *Proceedings of the the 1st ACM workshop on Continuous archival and retrieval of personal experiences (CARPE'04)*, New York, USA.

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

Jensen, F. V. (1996). *An Introduction to Bayesian Networks*. New York: Springer-Verlag.

Jensen, J., Lauritzen, S., and Olesen, K. (1990). Bayesian updating in recursive graphical models by local computations. *Computational Statistical Quarterly*, 4:269–282.

Jing, Y., Pavlovic, V., and Rehg, J. M. (2005). Efficient discriminative learning of bayesian network classifier via boosted augmented naive bayes. In *Proceedings of the 22nd International Conference on Machine Learning (ICML'05)*, pages 369–376, Bonn, Germany.

Jordan, M., Ghahramani, Z., Jaakkola, T., and Saul, L. (1999). An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233.

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

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

Jovanovic, N., op den Akker, R., and Nijholt, A. (2006b). A corpus for studying addressing behavior in multi-party dialogues. *Language Resources and Evaluation Journal*, 40(1):5–23.

Jurafsky, D., Shriberg, L., and Biasca, D. (1997). Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual, Draft 13. Technical Report TR-97-02, University of Colorado, The Institute of Cognitive Science, Boulder, CO.

Karaiskos, V. (2005). Reliability test for the AMI named entity annotation scheme (scenario meetings). Techical document. Available at http://mmm.idiap.ch/private/ami/annotation/NamedEntityReliabilityTest.pdf.

Katzenmaier, M., Stiefelhagen, R., and Schultz, T. (2004). Identifying the addressee in human-human-robot interactions based on head pose and speech. In *Proceedings of the 6th International Conference on Multimodal Interfaces (ICMI)*, State College, PA, USA.

Keizer, S. and op den Akker, R. (2006). Dialogue act recognition under uncertainty using Bayesian networks. *Natural Language Engineering*, pages 1–30. in Press.

Kendon, A. (1967). Some functions of gaze direction in social interaction. *Acta Psychologica*, 26:22–63.

Kerbrat-Orecchioni, C. (2004). Introducing polylogue. *Journal of Pragmatics*, 36(1):1–24.

Krippendorff, K. (1980). *Content Analysis: An Introduction to Its Methodology.* Beverly Hills, CA: Sage Publications.

Krippendorff, K. (2004a). *Content analysis: An Introduction to Its Methodology.* Thousand Oaks, CA: Sage, 2nd edition.

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

Lam, W. and Bacchus, F. (1994). Learning Bayesian belief networksan approach based on the MDL principle. *Computational Intelligence*, 10(4):269–293.

Landis, J. and Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174.

Langley, P., Iba, W., and Thompson, K. (1992). An analysis of Bayesian classifiers. In *Proceeding of the 10th National Conference on Artificial Intelligence*, San Jose, CA, USA.

Larrue, J. and Trognon, A. (1992). Organization of turn-taking and mechanisms for turn-taking repairs in a chaired meeting. *Journal of Pragmatics*, 19:177–196.

Lauritzen, S. and Spiegelhalter, D. (1988). Local computations with probabilities on graphical structures and their application to expert systems. *Journal of Royal Statistical Society (Series B)*, 50:157–224.

Lerner, G. H. (1991). On the syntax of sentences in progress. *Language In Society*, 20:441–458.

Lerner, G. H. (1993). Collectivities in action: Establishing the relevance of conjoined participation in conversation. *Text*, 13(2):213–245.

Lerner, G. H. (1995). Turn design and the organization of participation in instructional activities. *Discourse Processes*, 19(1):111–131.

Lerner, G. H. (1996a). Finding “face” in the preference structures of talk-in-interaction. *Social Psychology Quarterly*, 59(4):303–321.

Lerner, G. H. (1996b). On the place of linguistic resources in the organization of talk-in interaction: “Second person” reference in multi-party conversation. *Pragmatics*, 6(3):281–294.

Lerner, G. H. (2003). Selecting next speaker: The context-sensitive operation of a context-free organization. *Language in Society*, 32:177–201.

Lesch, S., Kleinbauer, T., and Alexandersson, J. (2005). Towards a decent recognition rate for the automatic classification of a multidimensional dialogue act tagset. In *4th Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, Edinburgh, Scotland.

Levinson, S. C. (1983). *Pragmatics.* Cambridge: Cambridge University Press.

Levinson, S. C. (1987). Putting linguistics on a proper footing: Explorations in goffman's participation framework. In Drew, P. and Wootton, A., editors, *Erving Goffman: Exploring the interaction order*, pages 161–227. Oxford: Polity Press.

Lisowska, A. (2003). Multimodal interface design for the multimodal meeting domain: Preliminary indications from a query analysis study. Technical Report 11, IM2.MDM.

Lunsford, R. and Oviatt, S. (2006). Human perception of intended addressee in multiparty meetings. In *Proceedings of 8th International Conference on Multimodal Interfaces (ICMI'06)*, Banff, Canada.

Manning, C. and Schutze, H. (1999). *Foundations of Statistical Natural Language Processing.* Cambridge, MA: The MIT Press.

McCowan, I., Bengio, S., Gatica-Perez, D., Lathoud, G., Monay, F., Moore, D., Wellner, P., and Bourlard, H. (2003). Modeling human interactions in meetings. In *Proceedings of 28th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Hong Kong.

Minka, T. P. (2001). *A family of algorithms for approximate Bayesian inference.* PhD thesis, MIT Media Lab.

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

Moran, T. P., L. Palen, S. H., Chiu, P., Kimber, D., Minneman, S., Melle, W. V., and Zellweger, P. (1997). I'll get that off the audio': A case study of salvaging multimedia meeting records. In *Proceedings of CHI*, Atlanta, Georgia, USA.

Murphy, K. and Weiss, Y. (2001). The Factored Frontier algorithm for approximate inference in DBNs. In *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence*, Seattle, WA.

Murphy, K. P. (2001). The Bayes Net toolbox for Matlab. *Computing Science and Statistics.*

Murphy, K. P. (2002). *Dynamic Bayesian Networks: Representation, Inference and Learning.* PhD thesis, UC Berkeley, Computer Science Division.

Nijholt, A. (2007). Google home: Experience, support and re-experience of social home activities. *Journal of Information Sciences, Special Issue on Ambient Intelligence, Elsevier, Amsterdam.* to Appear.

Nijholt, A., Rienks, R., Zwiers, J., and Reidsma, D. (2006). Online and off-line visualization of meeting information and meeting support. *The Visual Computer*, pages 1–12. To Appear.

O'Connell, D. C., Kowal, S., and Kaltenbacher, E. (1990). Turn-taking: a critical analysis of the research tradition. *Journal of Psycholinguistic Research*, 19(6):345–373.

Otsuka, K., Takemae, Y., Yamato, J., and 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 International Conference on Multimodal Interfaces (ICMI)*, pages 191–198, Trento, Italy.

Pallotta, V., Seretan, V., Ailomaa, M., Ghorbel, H., and Rajman, M. (2006). Query types for meeting information systems: assessing the role of argumentative structure in answering questions on meeting discussion records. Presented at MMAD: Modelling Meetings, Argumentation and Discourse Workshop, Liverpool, UK. Online.

Parker, K. (1988). Speaking turns in small group interaction: A context-sensitive event sequence model. *Journal of Personality and Social Psychology*, 54(6):965–971.

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

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

Pearl, J. (1982). Reverend bayes on inference engines: A distributed hierarchical approach. In *Proceedings of the National Conference on Artificial Intelligence*, Pittsburgh, PA.

Pearl, J. (1986). Fusion, propagation, and structuring in belief networks. *Artificial Intelligence*, 29(3):241–288.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Mateo, CA: Morgan Kaufmann Publishers.

Popescu-Belis, A. (2003). Dialogue act tagsets for meeting understanding: an abstraction based on the DAMSL, Switchboard and ICSI-MR tagsets. Technical Report IM2.MDM-09, ISSCO/TIM/ETI, University of Geneva. Version 1.2 (December 2004).

Popescu-Belis, A. (2005). Dialogue acts: One or more dimensions? ISSCO Working Paper 62, University of Geneva, Switzerland.

Post, W., Cremers, A., and Henkemans, O. (2004). A research environment for meeting behavior. In Nijholt, A., Nishida, T., Fruchter, R., and Rosenberg, D., editors, *Social Intelligence Design*, Enschede, The Netherlands.

Purver, M., Ehlen, P., and Niekrasz, J. (2006). Detecting action items in multi-party meetings: Annotation and initial experiments. In *Proceedings of the 3rd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms*, Washington, DC, USA.

Purver, M., Niekrasz, J., and Peters, S. (2005). Ontology-based multi-party meeting understanding. In *Proceedings of the CHI 2005 workshop The Virtuality Continuum Revisited*, Portland, OR.

Rehg, J. M., Murphy, K. P., and Fieguth, P. W. (1999). Vision-based speaker detection using Bayesian networks. In *Proceedings of Conference on Computer Vision and Pattern Recognition*, Ft. Collins, CO.

Reidsma, D., Hofs, D., and Jovanovic, N. (2005). Designing focused and efficient annotation tools. In Noldus, L., Grieco, F., Loijens, L., and Zimmerman, P., editors, *Measuring Behaviour, 5th International Conference on Methods and Techniques in Behavioral Research*, Wageningen, The Netherlands.

Rienks, R., Nijholt, A., and Barthelmess, P. (2007). Pro-active meeting assistants : Attention please! *AI and Society, The Journal of Human-Centred Systems.* to Appear.

Romano, N. C. and Nunamaker, J. F. (2001). Meeting analysis: findings from research and practice. In *Proceedings of the 34th Hawaii International Conference on System Sciences (HICSS-34).*

Russell, S. J. and Norvig, P. (2003). *Artificial Intelligence, A Modern Approach.* New Jersey: Prentice Hall, 2nd edition.

Sacks, H. (1992). *Lectures on Conversation. Volumes I and II*, volume I. Cambridge, MA: Basil Blackwell.

Sacks, H. and Schegloff, E. A. (1979). Two preferences in the organization of reference to persons in conversation and their interaction. In *G. Psathas, ed., Everyday language: studies in ethnomethodology*, pages 15–21. New York: Irvington.

Sacks, H., Schegloff, E. A., and Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, 50:696–735.

Schegloff, E. A. (1986). The routine as achievement. *Human Studies*, 9:111–151.

Schegloff, E. A. (1988). Goffman and the analysis of conversation. In Drew, P. and Wooton, A., editors, *Erving Goffman: Exploring the interaction order*, pages 89–135. Cambridge: Polity Press.

Schegloff, E. A. (1995). Parties and talking together: two ways in which numbers are significant in talk-in-interaction. In ten Have, P. and Psathas, G., editors, *Situated order: studies in the social organization of talk and embodied activities*, pages 31–42. Washington: University Press of America.

Schegloff, E. A. (1996). Some practices for referring to persons in talk-in-interaction: A partial sketch of a systematics. In Fox, B., editor, *Studies in Anaphora*, pages 437–485. Amsterdam: John Benjamins.

Schegloff, E. A. (2002). Opening sequencing. In Katz, J. and Aakhus, M., editors, *Perpetual Contact: Mobile communication, private talk, public performance*, pages 326–385. Cambridge: Cambridge University Press.

Schegloff, E. A. and Sacks, H. (1973). Opening up closings. *Semiotica*, 7(4):289–327.

Schegloff, E. A., Sacks, H., and Jefferson, G. (1977). The preference for self-correction in the organization of repair in conversation. *Language*, 53(2):361–382.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464.

Searle, J. (1969). *Speech Acts: An Essay in the Philosophy of Language.* Cambridge: Cambridge University Press.

Searle, J. (1975). Indirect speech acts. In Cole, P. and Morgan, J. L., editors, *Syntax and Semantics Volume 3: Speech Acts*, pages 55–82. New York: Academic Press.

Shachter, R. D., D'Ambrosio, B. D., and Favero, B. D. D. (1990). Symbolic probabilistic inference in belief networks. In *Proceeding of the 8th National Conference on Artificial Intelligence*, Boston, MA, USA.

Shaw, M. E. (1981). *Group dynamics: The psychology of small group behavior.* New York: McGraw-Hill.

Shriberg, E., Dhillon, R., Bhagat, S., J.Ang, and Carvey, H. (2004). The ICSI Meeting Recorder Dialog Act (MRDA) corpus. In *Proceedings of 5th SIGdial Workshop on Discourse and Dialogue*, pages 97–100, Boston, USA.

Shriberg, E., Stolcke, A., and Baron, D. (2001). Observations on overlap: findings and implications for automatic processing of multi-party conversation. In *Proceedings of7 th European Conference on Speech Communication and Technology (EUROSPEECH)*, Aalborg, Denmark.

Stasser, G. and Taylor, L. (1991). Speaking turns in face-to-face discussions. *Journal of Personality and Social Psychology*, 60:675–684.

Stiefelhagen, R., Yang, J., and Waibel, A. (2002). Modeling focus of attention for meeting indexing based on multiple cues. *IEEE Transactions on Neural Networks*, 13(4):928–938.

Stiefelhagen, R. and Zhu, J. (2002). Head orientation and gaze direction in meetings. In *Conference on Human Factors in Computing Systems (CHI2002)*, Minneapolis, MI, USA.

Traum, D. (2004). Issues in multi-party dialogues. In Dignum, F., editor, *Advances in Agent Communication*, pages 201–211. Springer-Verlag.

Traum, D. R. and Hinkelman, E. A. (1992). Conversation acts in task-oriented spoken dialogue. *Computational Intelligence*, 8(3):575–599.

Tucker, S. and Whittaker, S. (2005). Accessing multimodal meeting data: Systems, problems and possibilities. In Bengio, S. and Bourlard, H., editors, *Machine Learning for Multimodal Interaction, First International Workshop, MLMI 2004, Martigny, Switzerland, June 21-23, 2004, Revised Selected Papers*, volume 3361 of *Lecture Notes in Computer Science*, pages 1–11. Springer.

van Turnhout, K., Terken, J., Bakx, I., and Eggen, B. (2005). Identifying the intended addressee in mixed human-human and human-computer interaction from non-verbal features. In *Proceedings of 7th International Conference on Multimodal Interfaces (ICMI'05)*, Trento, Italy.

Vertegaal, R. (1998). *Look who is talking to whom.* PhD thesis, University of Twente, Enschede, The Netherlands.

Whittaker, S., Laban, R., and Tucker, S. (2006). Analysing meeting records: An ethnographic study and technological implications. In Renals, S. and Bengio, S., editors, *Machine Learning for Multimodal Interaction, Second International Workshop, MLMI 2005, Edinburgh, UK, July, 2005, Revised Selected Papers*, volume 3869 of *Lecture Notes in Computer Science*, pages 101–113. Springer-Verlag.

Witten, I. H. and Frank, E. (2000). *Data mining: Practical machine learning tools and techniques with Java implementations.* San Francisco: Morgan Kaufmann.

Zhang, N. L. and Poole, D. (1994). A simple approach to Bayesian network computations. In *Proceedings of the Tenth Canadian Conference on Artificial Intelligence*, Banff, Alberta, Canada.

Zimmermann, M., Liu, Y., Shriberg, E., and Stolcke, A. (2006a). Toward joint segmentation and classification of dialog acts in multiparty meetings. In Renals, S. and Bengio, S., editors, *Machine Learning for Multimodal Interaction, Second International Workshop, MLMI 2005, Revised Selected Papers*, volume 3869 of *Lecture Notes in Computer Science*, pages 187–193. Springer-Verlag.

Zimmermann, M., Stolcke, A., and Shriberg, E. (2006b). Joint segmentation and classification of dialog acts in multiparty meetings. In *Proceedings of the 31st International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Toulouse, France.

SIKS Dissertation Series

1998-01 Johan van den Akker (CWI), DEGAS - An Active, Temporal Database of Autonomous Objects

1998–02 Floris Wiesman (UM), Information Retrieval by Graphically Browsing Meta-Information

1998-03 Ans Steuten (TUD), A Contribution to the Linguistic Analysis of Business Conversations within the Language/Action Perspective

1998-04 Dennis Breuker (UM), Memory versus Search in Games

1998-05 E.W.Oskamp (RUL), Computerondersteuning bij Straftoemeting

1999-01 Mark Sloof (VU), Physiology of Quality Change Modelling; Automated modelling of Quality Change of Agricultural Products

1999-02 Rob Potharst (EUR), Classification using decision trees and neural nets

1999-03 Don Beal (UM), The Nature of Minimax Search

1999-04 Jacques Penders (UM), The practical Art of Moving Physical Objects

1999-05 Aldo de Moor (KUB), Empowering Communities: A Method for the Legitimate User-Driven Specification of Network Information Systems

1999-06 Niek J.E. Wijngaards (VU), Re-design of compositional systems

1999-07 David Spelt (UT), Verification support for object database design

1999-08 Jacques H.J. Lenting (UM), Informed Gambling: Conception and Analysis of a Multi-Agent Mechanism for Discrete Reallocation.

2000-01 Frank Niessink (VU), Perspectives on Improving Software Maintenance

2000-02 Koen Holtman (TUE), Prototyping of CMS Storage Management

2000-03 Carolien M.T. Metselaar (UVA), Sociaal-organisatorische gevolgen van kennistechnologie; een procesbenadering en actorperspectief.

2000-04 Geert de Haan (VU), ETAG, A Formal Model of Competence Knowledge for User Interface Design

2000-05 Ruud van der Pol (UM), Knowledge-based Query Formulation in Information Retrieval.

2000-06 Rogier van Eijk (UU), Programming Languages for Agent Communication

2000-07 Niels Peek (UU), Decision-theoretic Planning of Clinical Patient Management

2000-08 Veerle Coup (EUR), Sensitivity Analyis of Decision-Theoretic Networks

2000-09 Florian Waas (CWI), Principles of Probabilistic Query Optimization

2000-10 Niels Nes (CWI), Image Database Management System Design Considerations, Algorithms and Architecture

2000-11 Jonas Karlsson (CWI), Scalable Distributed Data Structures for Database Management

2001-01 Silja Renooij (UU), Qualitative Approaches to Quantifying Probabilistic Networks

2001-02 Koen Hindriks (UU), Agent Programming Languages: Programming with Mental Models

2001-03 Maarten van Someren (UvA), Learning as problem solving

2001-04 Evgueni Smirnov (UM), Conjunctive and Disjunctive Version Spaces with Instance-Based Boundary Sets

2001-05 Jacco van Ossenbruggen (VU), Processing Structured Hypermedia: A Matter of Style

2001-06 Martijn van Welie (VU), Task-based User Interface Design

2001-07 Bastiaan Schonhage (VU), Diva: Architectural Perspectives on Information Visualization

2001-08 Pascal van Eck (VU), A Compositional Semantic Structure for Multi-Agent Systems Dynamics.

2001-09 Pieter Jan 't Hoen (RUL), Towards Distributed Development of Large Object-Oriented Models, Views of Packages as Classes

2001-10 Maarten Sierhuis (UvA), Modeling and Simulating Work Practice BRAHMS: a multiagent modeling and simulation language for work practice analysis and design

2001-11 Tom M. van Engers (VUA), Knowledge Management: The Role of Mental Models in Business Systems Design

2002-01 Nico Lassing (VU), Architecture-Level Modifiability Analysis

2002-02 Roelof van Zwol (UT), Modelling and searching web-based document collections

2002-03 Henk Ernst Blok (UT), Database Optimization Aspects for Information Retrieval

2002-04 Juan Roberto Castelo Valdueza (UU), The Discrete Acyclic Digraph Markov Model in Data Mining

2002-05 Radu Serban (VU), The Private Cyberspace Modeling Electronic Environments inhabited by Privacy-concerned Agents

2002-06 Laurens Mommers (UL), Applied legal epistemology; Building a knowledge-based ontology of the legal domain

2002-07 Peter Boncz (CWI), Monet: A Next-Generation DBMS Kernel For Query-Intensive Applications

2002-08 Jaap Gordijn (VU), Value Based Requirements Engineering: Exploring Innovative E-Commerce Ideas

2002-09 Willem-Jan van den Heuvel (KUB), Integrating Modern Business Applications with Objectified Legacy Systems

2002-10 Brian Sheppard (UM), Towards Perfect Play of Scrabble

2002-11 Wouter C.A. Wijngaards (VU), Agent Based Modelling of Dynamics: Biological and Organisational Applications

2002-12 Albrecht Schmidt (Uva), Processing XML in Database Systems

2002-13 Hongjing Wu (TUE), A Reference Architecture for Adaptive Hypermedia Applications

2002-14 Wieke de Vries (UU), Agent Interaction: Abstract Approaches to Modelling, Programming and Verifying Multi-Agent Systems

2002-15 Rik Eshuis (UT), Semantics and Verification of UML Activity Diagrams for Workflow Modelling

2002-16 Pieter van Langen (VU), The Anatomy of Design: Foundations, Models and Applications

2002-17 Stefan Manegold (UVA), Understanding, Modeling, and Improving Main-Memory Database Performance

2003-01 Heiner Stuckenschmidt (VU), Ontology-Based Information Sharing in Weakly Structured Environments

2003-02 Jan Broersen (VU), Modal Action Logics for Reasoning About Reactive Systems

2003-03 Martijn Schuemie (TUD), Human-Computer Interaction and Presence in Virtual Reality Exposure Therapy

2003-04 Milan Petković (UT), Content-Based Video Retrieval Supported by Database Technology

2003-05 Jos Lehmann (UVA), Causation in Artificial Intelligence and Law - A modelling approach

2003-06 Boris van Schooten (UT), Development and specification of virtual environments

2003-07 Machiel Jansen (UvA), Formal Explorations of Knowledge Intensive Tasks

2003-08 Yongping Ran (UM), Repair Based Scheduling

2003-09 Rens Kortmann (UM), The resolution of visually guided behaviour

2003-10 Andreas Lincke (UvT), Electronic Business Negotiation: Some experimental studies on the interaction between medium, innovation context and culture

2003-11 Simon Keizer (UT), Reasoning under Uncertainty in Natural Language Dialogue using Bayesian Networks

2003-12 Roeland Ordelman (UT), Dutch speech recognition in multimedia information retrieval

2003-13 Jeroen Donkers (UM), Nosce Hostem - Searching with Opponent Models

2003-14 Stijn Hoppenbrouwers (KUN), Freezing Language: Conceptualisation Processes across ICT-Supported Organisations

2003-15 Mathijs de Weerdt (TUD), Plan Merging in Multi-Agent Systems

2003-16 Menzo Windhouwer (CWI), Feature Grammar Systems - Incremental Maintenance of Indexes to Digital Media Warehouses

2003-17 David Jansen (UT), Extensions of Statecharts with Probability, Time, and Stochastic Timing

2003-18 Levente Kocsis (UM), Learning Search Decisions

2004-01 Virginia Dignum (UU), A Model for Organizational Interaction: Based on Agents, Founded in Logic

2004-02 Lai Xu (UvT), Monitoring Multi-party Contracts for E-business

2004-03 Perry Groot (VU), A Theoretical and Empirical Analysis of Approximation in Symbolic Problem Solving

2004-04 Chris van Aart (UVA), Organizational Principles for Multi-Agent Architectures

2004-05 Viara Popova (EUR), Knowledge discovery and monotonicity

2004-06 Bart-Jan Hommes (TUD), The Evaluation of Business Process Modeling Techniques

2004-07 Elise Boltjes (UM), Voorbeeldig onderwijs; voorbeeldgestuurd onderwijs, een opstap naar abstract denken, vooral voor meisjes

2004-08 Joop Verbeek(UM), Politie en de Nieuwe Internationale Informatiemarkt, Grensregionale politile gegevensuitwisseling en digitale expertise

2004-09 Martin Caminada (VU), For the Sake of the Argument; explorations into argument-based reasoning

2004-10 Suzanne Kabel (UVA), Knowledge-rich indexing of learning-objects

2004-11 Michel Klein (VU), Change Management for Distributed Ontologies

2004-12 The Duy Bui (UT), Creating emotions and facial expressions for embodied agents

2004-13 Wojciech Jamroga (UT), Using Multiple Models of Reality: On Agents who Know how to Play

2004-14 Paul Harrenstein (UU), Logic in Conflict. Logical Explorations in Strategic Equilibrium

2004-15 Arno Knobbe (UU), Multi-Relational Data Mining

2004-16 Federico Divina (VU), Hybrid Genetic Relational Search for Inductive Learning

2004-17 Mark Winands (UM), Informed Search in Complex Games

2004-18 Vania Bessa Machado (UvA), Supporting the Construction of Qualitative Knowledge Models

2004-19 Thijs Westerveld (UT), Using generative probabilistic models for multimedia retrieval

2004-20 Madelon Evers (Nyenrode), Learning from Design: facilitating multidisciplinary design teams

2005-01 Floor Verdenius (UVA), Methodological Aspects of Designing Induction-Based Applications

2005-02 Erik van der Werf (UM), AI techniques for the game of Go

2005-03 Franc Grootjen (RUN), A Pragmatic Approach to the Conceptualisation of Language

2005-04 Nirvana Meratnia (UT), Towards Database Support for Moving Object data

2005-05 Gabriel Infante-Lopez (UVA), Two-Level Probabilistic Grammars for Natural Language Parsing

2005-06 Pieter Spronck (UM), Adaptive Game AI

2005-07 Flavius Frasincar (TUE), Hypermedia Presentation Generation for Semantic Web Information Systems

2005-08 Richard Vdovjak (TUE), A Model-driven Approach for Building Distributed Ontology-based Web Applications

2005-09 Jeen Broekstra (VU), Storage, Querying and Inferencing for Semantic Web Languages

2005-10 Anders Bouwer (UVA), Explaining Behaviour: Using Qualitative Simulation in Interactive Learning Environments

2005-11 Elth Ogston (VU), Agent Based Matchmaking and Clustering - A Decentralized Approach to Search

2005-12 Csaba Boer (EUR), Distributed Simulation in Industry

2005-13 Fred Hamburg (UL), Een Computermodel voor het Ondersteunen van Euthanasiebeslissingen

2005-14 Borys Omelayenko (VU), Web-Service configuration on the Semantic Web; Exploring how semantics meets pragmatics

2005-15 Tibor Bosse (VU), Analysis of the Dynamics of Cognitive Processes

2005-16 Joris Graaumans (UU), Usability of XML Query Languages

2005-17 Boris Shishkov (TUD), Software Specification Based on Reusable Business Components

2005-18 Danielle Sent (UU), Test-selection strategies for probabilistic networks

2005-19 Michel van Dartel (UM), Situated Representation

2005-20 Cristina Coteanu (UL), Cyber Consumer Law, State of the Art and Perspectives

2005-21 Wijnand Derks (UT), Improving Concurrency and Recovery in Database Systems by Exploiting Application Semantics

2006-01 Samuil Angelov (TUE), Foundations of B2B Electronic Contracting

2006-02 Cristina Chisalita (VU), Contextual issues in the design and use of information technology in organizations

2006-03 Noor Christoph (UVA), The role of metacognitive skills in learning to solve problems

2006-04 Marta Sabou (VU), Building Web Service Ontologies

2006-05 Cees Pierik (UU), Validation Techniques for Object-Oriented Proof Outlines

2006-06 Ziv Baida (VU), Software-aided Service Bundling - Intelligent Methods & Tools for Graphical Service Modeling

2006-07 Marko Smiljanić (UT), XML schema matching – balancing efficiency and effectiveness by means of clustering

2006-08 Eelco Herder (UT), Forward, Back and Home Again - Analyzing User Behavior on the Web

2006-09 Mohamed Wahdan (UM), Automatic Formulation of the Auditor's Opinion

2006-10 Ronny Siebes (VU), Semantic Routing in Peer-to-Peer Systems

2006-11 Joeri van Ruth (UT), Flattening Queries over Nested Data Types

2006-12 Bert Bongers (VU), Interactivation - Towards an e-cology of people, our technological environment, and the arts

2006-13 Henk-Jan Lebbink (UU), Dialogue and Decision Games for Information Exchanging Agents

2006-14 Johan Hoorn (VU), Software Requirements: Update, Upgrade, Redesign - towards a Theory of Requirements Change

2006-15 Rainer Malik (UU), CONAN: Text Mining in the Biomedical Domain

2006-16 Carsten Riggelsen (UU), Approximation Methods for Efficient Learning of Bayesian Networks

2006-17 Stacey Nagata (UU), User Assistance for Multitasking with Interruptions on a Mobile Device

2006-18 Valentin Zhizhkun (UVA), Graph transformation for Natural Language Processing

2006-19 Birna van Riemsdijk (UU), Cognitive Agent Programming: A Semantic Approach

2006-20 Marina Velikova (UvT), Monotone models for prediction in data mining

2006-21 Bas van Gils (RUN), Aptness on the Web

2006-22 Paul de Vrieze (RUN), Fundaments of Adaptive Personalisation

2006-23 Ion Juvina (UU), Development of Cognitive Model for Navigating on the Web

2006-24 Laura Hollink (VU), Semantic Annotation for Retrieval of Visual Resources

2006-25 Madalina Drugan (UU), Conditional log-likelihood MDL and Evolutionary MCMC

2006-26 Vojkan Mihajlović (UT), Score region algebra: a flexible framework for structured information retrieval

2006-27 Stefano Bocconi (CWI), Vox Populi: generating video documentaries from semantically annotated media repositories

2006-28 Borkur Sigurbjornsson (UVA), Focused Information Access using XML Element Retrieval

2007-01 Kees Leune (UvT), Access Control and Service-Oriented Architectures

2007-02 Wouter Teepe (RUG), Reconciling Information Exchange and Confidentiality: A Formal Approach

2007-03 Peter Mika (VU), Social Networks and the Semantic Web

2007-04 Jurriaan van Diggelen (UU), Achieving Semantic Interoperability in Multi-agent Systems: a dialogue-based approach

2007-05 Bart Schermer (UL), Software Agents, Surveillance, and the Right to Privacy: a Legislative Framework for Agent-enabled Surveillance

2007-06 Gilad Mishne (UVA), Applied Text Analytics for Blogs

2007-07 Nataša Jovanović (UT), To Whom It May Concern - Addressee Identification in Face-to-Face Meetings

Summary

This thesis is concerned with automatic addressee identification in face-to-face meetings. A grate number of meetings are organized everyday around the world to exchange ideas, share information, negotiate alternatives and make decisions. Despite their importance and prevalence there is a general observation that meetings are neither as productive nor as efficient as we would like them to be. To overcome these problems, technology to support meetings has being developed. With advances in multimedia technologies, it has become feasible to record all aspects of interaction taking place in meetings and thus to enhance traditional ways of documenting meetings: meeting minutes and personal notes. To make benefits out of meeting recordings, the recorded content needs to be automatically analyzed: relevant information needs to be extracted from the meeting content and the extracted information needs to be structured in such a way as to enable an efficient access to meeting data. Recently, research has begun to develop in that direction establishing a new research area: automatic meeting analysis. Since meetings are a domain where multi-party dialogues take place, automatic analysis and understanding of meeting discussions comprises not only identification of who speaks, what is said or what type of communicative act the speaker performs with his utterance but also *to whom the speaker addresses that act.*

To model addressing, we need to understand the basic processes that underlie this activity. The first part of the thesis is thus devoted to gaining theoretical insights into addressing based on the outcomes of the research in conversational and interaction analysis. Addressing is communicated through various communication channels such as speech, gaze, gestures and postures. In many cases, addressing is accomplished not explicitly but tacitly. In the latter case, the content of what has been said and various features of conversational context play the major role. The multi-modal nature of addressing as well as its context sensitivity pose challenges not only for computational systems but also for humans in determining who is being addressed in a particular situation. This thesis investigates the difficulties of the task of addressee identification for both humans and systems.

To develop computational models for addressee prediction, a collection of audio and video meeting recordings annotated with addressee information as well as with phenom-

ena related to addressing is required (e.g. where participants are looking, what type of communicative act the speaker is performing). The second part of the thesis describes two meeting corpora employed in our study. The corpora were developed using annotation schemas designed in the interaction between data observed and theoretical insights into addressing obtained from the literature. To assess the credibility of the annotated data for drawing research results relying on them, the thesis provides an exhaustive reliability analysis of the annotation schemas. A detailed investigation of the problems human observers had in determining who is being addressed by the speaker shows that it is intrinsically difficult to distinguish between group and individual addressing. This certainly brings an additional challenge in modeling this process automatically.

The third part of the thesis deals with the development of a computational model for automatic addressee identification. As a computational framework, we use Bayesian Networks. Features employed to model addressing are obtained from speech and gaze communication channels as well as from conversational and meeting contexts. Conversational context features, which relate to conversational history, are shown to be the most valuable for addressee prediction; they dominate over utterance and gaze features. We observe that utterance features, which contain a set of lexical features extracted from the utterance content as well as the type of dialogue act performed with the utterance, are the most uninformative cues for determining the intended addressee. Furthermore, knowing who is looking at whom in a meetings does not provide sufficient information for determining who is talking to whom during the meeting, although one would intuitively expect that it would. Meeting context, which comprises various types of activities participants perform in a meeting, does not add valuable information for addressee identification to utterance, gaze and contextual features. By combining all these four types of features addressees are predicted with the highest accuracy.

Evaluation of several static Bayesian Network classifiers indicates that Bayesian Networks are effective computational models for the task of addressee prediction. Among all evaluated models, augmented Naive Bayes classifiers show the best performances over all feature sets. Similar to human annotators, Bayesian Network addressee classifiers "observe" addressing in a particular situation using a set of defining features and classify the observed process into several categories denoting the intended addressee: a particular individual or a group. In that sense, Bayesian Network classifiers can be considered as automatic annotators. The present study reveals that Bayesian annotators behave in a similar way to human annotators regarding the type confusion between addressee values.

As we rely on hand-annotated values for the input features instead of automatically computed values based on measurements, the present study aims to investigate the upper bounds for the results that could be obtained in a more realistic, fully-automatic scenario. As the first step in the automation of addressee detection process, we propose a Dynamic Bayesian Network model. The evaluation of performances of addressee classifiers relying on fully automatic features remains one of the tasks for future research.

Samenvatting

Dit proefschrift gaat over het automatisch identificeren van die deelnemer(s) die door de spreker in een vergadering word(t)(en) aangesproken. Dagelijks worden over de hele wereld een groot aantal vergaderingen gehouden, vergaderingen waarin ideeën worden uitgewisseld en waarin besluiten worden genomen. Veel vergaderingen zijn niet zo productief en efficiënt als we zouden willen. Om vergaderingen productiever en efficiënter te maken wordt speciale ondersteunende technologie ontwikkeld. Door ontwikkeling van multimediatechnieken ontstaan nieuwe mogelijkheden voor het automatiseren van het samenvatten van vergaderingen, zowel in video- als in audio- als in tekstformaat. Daarvoor is het nodig automatisch te bepalen waarover de deelnemers aan de vergadering het hebben en welke activiteiten er tijdens de vergadering worden uitgevoerd, welke beslissingen zijn genomen en hoe deze beslissingen zijn genomen. De multimodale gegevens die het resultaat zijn van een extractie- en herkenningsproces moeten op zodanige wijze gestructureerd worden opgeslagen dat het mogelijk is op efficiënte wijze de gewenste gegevens op te zoeken. Een nieuw onderzoeksgebied vormt de context van het onderzoek waar dit proefschrift toe bijdraagt: automatische analyse van vergaderingen. Omdat in vergaderingen meerdere deelnemers betrokken zijn, moet niet alleen worden vastgesteld wie er spreekt en waarover gesproken wordt, maar ook tegen wie de spreker spreekt, ofwel *wie de geadresseerde van de spreker is.*

Om het fenomeen van adressering te modelleren, moeten we de basisprocessen van deze activiteit begrijpen. Het eerste deel van dit proefschrift is daarom gewijd aan het verkrijgen van een theoretisch kader. Daartoe baseren we ons op bestaande theorieën uit de conversatie-analyse en de interactie-analyse van menselijk gedrag. Adresseren is een activiteit die gerealiseerd wordt door een scala van verbale en niet-verbale gedragingen: spraak, kijkgedrag, gebaren en houdingen. In veel gevallen wordt adressering niet expliciet maar impliciet gerealiseerd. In het laatste geval spelen diverse eigenschappen van de context van het gesprek de belangrijkste rol bij het bepalen van de geadresseerde. Zowel het multimodale karakter van het adresseergedrag als de context-afhankelijkheid ervan vormen een uitdaging voor het identificeren van de geadresseerde. En dat geldt zowel voor het detecteren door een machine als door de mens. Dit proefschrift bespreekt de moeilijkheden van deze taak: het vinden van de geadresseerde door de mens en door de machine, en geeft aanzetten tot het oplossen ervan.

Ten behoeve van het modelleren van adressering is een collectie van audio- en video-opnames van vergaderingen geannoteerd met adresseringsinformatie. Van iedere spreekactiviteit (*speech act*) is door mensen aangegeven wie de in die taalhandeling geadresseerde is: een individu of een (sub)groep van aanwezigen. Tevens zijn een aantal voor adresseringdetectie relevante kenmerken van gedragingen en context geannoteerd. Het tweede deel van het proefschrift beschrijft twee collecties (corpora) van vergaderingen. Deze corpora werden ontwikkeld met behulp van annotatieschema's die de annotatoren voorschrijven hoe ze de verschillende observeerbare fenomenen moeten segmenteren en van een label moeten voorzien. Deze schema's zijn gebaseerd op de theoretische inzichten die uiteengezet zijn in het eerste deel. Er is een uitvoerige *reliability-analyse* uitgevoerd op de diverse annotaties om inzicht te krijgen in de mate waarin verschillende annotatoren die eenzelfde vergadering observeren de fenomenen in deze vergadering segmenteren en labelen. Een gedetailleerde analyse van de verschillen tussen annotatoren laat zien dat het in sommige gevallen moeilijk is om te zeggen of een spreker een individu dan wel de hele groep adresseert. Dit vormt een extra uitdaging voor het automatiseren van dit proces.

Het derde deel van dit proefschrift gaat over het ontwikkelen van een computationeel model voor het automatisch identificeren van de geadresseerde van de spreker. Daarvoor gebruiken we Bayesiaanse Netwerken, waarvan de knopen staan voor de relevante eigenschappen van spraak, kijkgedrag en van kenmerken van de gesprekscontext. De aan een taalhandeling voorafgaande taalhandelingen blijken de meest informatieve indicatoren te zijn voor het bepalen van de geadresseerde van de spreker. Ze zijn belangrijker dan lexicale kenmerken van de taalhandeling zelf en het kijkgedrag van de spreker en de andere deelnemers op het moment dat die handeling wordt uitgevoerd. Het kijkgedrag van de andere deelnemers is het minst informatief. Verder is het zo dat kijkgedrag (wie kijkt naar wie) niet voldoende is om te bepalen tot wie de spreker zich richt. Wat men misschien wel zou verwachten. Ook informatie over de soort vergaderactiviteit waarbinnen de taalhandeling wordt uitgevoerd is niet voldoende voor detectie van de geadresseerde. Maar voegen we al deze soorten indicatoren samen, dan kunnen we redelijk goed voorspellen wie er geadresseerd wordt.

Evaluatie van diverse varianten van Bayesiaanse Netwerken, getraind met (een deel van) de geannoteerde data, toont aan dat deze modellen geschikt zijn voor deze taak. Onder de geëvalueerde modellen leveren de *augmented Naive Bayesian Networks* de beste resultaten gemeten naar het percentage van juist geadresseerde acties. Ze geven het hoogste percentage correcte resultaten wanneer we testen op een ongezien deel van het geannoteerde corpus. Evenals menselijke annotatoren annoteren deze netwerken de taalhandelingen met een adresseerlabel. In die zin kunnen we de ontwikkelde classificatiesystemen dus beschouwen als automatische annotatoren. Het blijkt dat ze zich op vergelijkbare wijze gedragen als menselijke annotatoren, wanneer we kijken naar de soort fouten die ze maken. Omdat de tests uitgevoerd zijn op handgeannoteerde data, leveren de resultaten van de evaluaties bovengrenzen van de accuratesse die automatische systemen voor adresseringsclassificaties kunnen bereiken. Wanneer de voor deze predictie gebruikte invoerwaarden van het model ook volledig automatisch bepaald worden, zullen door fouten bij deze automatische bepaling lagere resultaten bereikt worden. Als eerste stap naar een verdere automatisering van geadresseerdenidentificatie zijn Dynamische Bayesiaanse Netwerken gemaakt en getest. Het ontwikkelen en evalueren van een systeem dat volledig automatisch dit werk doet, blijft werk voor de toekomst.

Acknowledgments

Line after line, page after page and here we are. I finally get the unique opportunity to express my gratitude to all those people who contributed to the completion of the thesis that lies before you.

First of all, I would like to thank my promoter Anton Nijholt for giving me an opportunity to carry out my PhD project in the Human Media Interaction (HMI) group. I very much appreciate the balance he has struck between giving me the freedom to conduct the research in my way and guiding me in the right direction through my PhD journey. Regarding my research work, I am most in debt to my daily supervisor Rieks op den Akker for his everlasting support and encouragement as well as for his positivism about my research. No matter how busy he was, Rieks has always found time for me. Being particularly interested in my research topic, Rieks was helpful in each stage of my work. Many ideas implemented in this thesis were inspired by our endless discussions on addressing. What I will miss the most about Rieks is his optimism, enthusiasm and above everything his sarcastic sense of humor.

I would like to express my gratitude to all my committee members for taking the time to read my thesis and to evaluate my work. I would also like to thank them for useful comments that have improved the quality of this thesis.

I am very grateful to Mannes Poel for his proofreading of the experimental part of the thesis, valuable discussions about Bayesian Networks and critical comments on my work. I am also thankful to Mariët Theune for showing interest in my research and for providing me with insightful comments on the introductory chapter. Many thanks to Lynn Packwood for correcting my English and thus making a reader's life much easier.

I was lucky to get an opportunity to do my PhD in the scope of two international projects: M4 and AMI. Working in collaboration with several universities and research institutes from Europe and USA, was, in the first place, a unique professional experience. I would like to acknowledge the M4 and AMI research teams for the great cooperation we had in the past four years.

It has been a great pleasure working with all my former and current colleagues in the HMI group. Although they all contributed in one way or another to the completion of this

thesis, several of them deserve special credits. I feel short of words to express my gratitude to Dennis Reidsma who has been generously providing me with technical and scientific support all these years. Without him showing an interest in NXT a few years ago, this thesis would certainly have a different content and title than it has today. Just to confirm my hypothesis, Dennis actually came up with the original title of the thesis, putting in that way a final stamp to his contribution. Many special thanks to my colleague, friend and house mate Yulia for all encouragement and support she has been providing me in the last two years. I am also thankful to my officemate Wauter Bosma for the friendly atmosphere we had in our office accompanied with interesting discussions regarding both work and non-work related issues. Many of my troubles with the Dutch language were solved with his kind help. Special thanks to my friend Ronald Poppe (aka Ronaldiño) who has always been trying to cheer me up even when there was no need for that. I am especially grateful to him for kindly helping me with the design of the thesis' cover. Many joyful discussions I had with Martijn van Otterlo helped me to get to know the Dutch culture and lifestyle better. Special thanks to him for sharing with me his knowledge and experience in machine learning. Finally, many thanks to the female group of PhD students at HMI: Olga, Andreea, Claudia and Yulia (previously mentioned).

Many of the technical troubles would have never been solved without help of Hendri Hondorp with whom I have learned that hacking can sometimes be a fun. I am sure that a reader would not like to know how to hack WinEdit to make it work (again!). Regarding the technical support, I am particularly thankful to Dennis Hofs for his valuable assistance during the implementation stage. For helping me with the tedious paperwork, I am very much grateful to our secretaries Charlotte and Alice. Special thanks to Charlotte and Rieks for translating the thesis summary in Dutch.

Living abroad would have been much more difficult without all my friends I have met in the Netherlands. Without them this journey would have been much more stressful and much less enjoyable. Since my first days of arriving in the Netherlands, Zlatko was the one who tried to add a bit of the light to the darkness I had been faced with (it was December and it was different). My deepest thankfulness to him for being always there for me. Many special and heartfelt thanks to a group of my "NL" friends with whom I shared the most important moments in our lives in the past four years: Tanja, Ana, Monija, Mira (Mladja&Gavra), Danijela, Darija and Mladen. Not least, I thank to Dule, Jelena M., Dragančе, Vojkan&Sofka, Stanislav&Vania&Twins, Nicky, Marko&family, Maja, Boris&Jelena, Zoki&Tanja, Saša, Boki, Raša, Ljuba and many others for the great time we had together. I have to send an additional special thank to my *paranimfen* Tanja and Ana for going with me through the most difficult year in my life so far as well as to Dule for being a reliable source of information of all kinds.

A big thank goes to my NYC dreamer for giving me an opportunity to escape, at least for a moment, in a dreamland where nothing but smiles exist. Knowing that there is such a place, gave me the necessary strength and inspiration to finalize this thesis.

With all my heart I thank my friends in Serbia - Ivana, Živan, Tanja, Gigi, Maša, Devi and Nena - for all their love and encouragement. I'm happy to know that in spite of the distance the strength of our friendships has stayed unchanged. On the other hand,

I “blame” them for having the constant feeling that, in the last four years, I have been living somewhere in between Serbia and the Netherlands.

One of the most beautiful things that has happened in my life so far is the friendship with my best friend Viktor. He has contributed to a large extent to my personal and above everything to my professional development. Working together for several years, we have become a promising team that could have coped with any problem though not necessarily in the most efficient and the most optimal way. Viktor has always believed in me more than I have ever believed in myself. His everlasting and sometimes too ambitious support resulted in having this thesis written. Here I have to stop writing about Viktor’s importance in all aspects of my life otherwise this chapter may end up with much more pages then it is common. All my gratitude for everything he did for me in the last four years can be summarized in the following words: Viktore, we made it!

I can’t not mention a tone of love and care I have received from my aunt Dara. Her support in the last days of writing this thesis was precious. I am also grateful to all my other relatives who have never doubted that I will come to the point where I am now. My greatest thanks, however, go to my grandmothers who have just simply loved me in a way that only grandmothers can love.

Last but not least, I tender my heartfelt acknowledgment to my parents Dragan and Caca and to my brother Ljubiša. Their endless love, support and belief were my “steering force” throughout whole my life. The love and admiration I feel for them is indescribable.

> *Dragi moji, beskrajno sam vam zahvalna za svu vašu ljubav i veru u mene. Hvala vam što ste mi dozvolili da verujem da su sve moje dosadašnje odluke bile one prave i nepogrešive. Srećna sam jer znam da će na svakoj sledećoj raskrsnici života svetleti zeleno svetlo koje znači “Napred” ali isto tako da će iza svakog ugla postojati skretanje u ulicu sigurnosti. Volim vas do neba!!!*

The End or To Be Continued?

Nataša Jovanović,
Enschede,
March 2007