Meetings in Smart Environments
Implications of Progressing Technologies

Rutger Rienks

April 10, 2007

Contents

Chapter 1

Introduction

1.1 Human Computing and Smart Environments

In a recent article[1] that predicts the future in 2015, an apparatus called the 'Perkomat' is introduced: a coffee machine that monitors business meetings and brews automatically when a meeting is stagnating. Forecasts predict that it can lead to productivity gains of 10% for some groups. An enhanced version, at that time under construction, will even pass frozen cookie dough through an oven when sensors detect a deterioration in interpersonal relationships during a meeting. These and other views of the future where humans are surrounded by interfaces embedded in all kinds of objects in the environment and being responsive to human presence have been an inspiration for scientist in the last decades. Advancing research in the area of human-computer interaction, smart environments, multi-modal interaction, ambient intelligence and ubiquitous computing nowadays is converging into the dawning era of human computing (Pantic et al., 2006). According to Jaimes et al. (2006) computing is nowadays getting towards one of its most exciting moments in history and is starting to play an essential role in supporting human activities.

Human computing inherits the complexity related to software engineering and system integration whilst embedding the human in the loop. And it inherits the difficulties of understanding and modelling human-human and human-computer interaction in the context of a changing environment (Clancey, 1997). Emerging systems are expected to be more and more of a different nature and to leap beyond the traditional productivity-oriented workplace technologies in which performance is the key objective. Interfaces for human computing go beyond keyboard and mouse and the interaction will no longer be determined by the predefined task and expected users alone. Back in 1991 it was already predicted that the applications of the 21st century would encompass leisure,

[1] 'Worldco in Lockdown after Riskometer tripped' was written by Jean Carletta as a preparation for an international brainstorm session aiming to explore directions for future scientific research

play, culture and art. They were expected to be increasingly implicit and more and more interweaved with the fabric of daily life (Weiser, 1991).

Compared to traditional systems, the following trends can be identified:

- **New sensing possibilities** The development and improvement of sensing technologies allows for the design of a broader spectrum of computer interfaces fostering the inclusion of so-called 'natural' interfaces that are created to enable 'intuitive' interaction. The area of language understanding has moved from speech recognition to human writing and human gesture recognition. Also developments in tactual interfaces such as haptic devices, biometric sensors, and in the recognition of non-speech back-channelling sounds (cf. Yngve (1970)) such as laughter and clapping is taking ground (Turk and Kolsch, 2004).

- **Shift in initiative** Traditional systems are responsive by nature, and dialogues with the user are guided by prescribed scenarios directly related to the goal of the system and its residing grammars. Nowadays, Human-Computer Interaction is becoming more and more a mixed-initiative in which humans and computers are engaged in less restricted dialogues. And looking ahead, one sees pro-active and context sensing systems appearing that suggest (or even perform) theater plays, and fulfill the role of social actor in an augmented-reality environment (Ju and Leifer, pear).

- **Diversifying physical interfaces** The physical forms of interfaces are diversifying (Benford et al., 2005). On the one hand the size of immersive displays and interactive billboards is growing, whereas on the other hand interfaces are becoming increasingly smaller and embedded in wearables (Tan and Pentland, 1997). Since the 1990's the domain of wearable computers took off due to developments in low power sensors, networking and component size issues (Thorpe, 1998). With the increase of all sorts of sensor networks and bandwidth, it nowadays has become possible to (collaboratively) interact remotely with each other and with applications.

- **Shift in application purpose** Whereas traditional systems are in general task-based, new applications are more and more focussed on the user's everyday dynamics (Benford et al., 2005). Along with this trend the concept of User Experience (UX) came along. For some current applications the task is no longer the goal, but rather the interaction itself (e.g. Reidsma et al. (2006)). Aspects such as beauty, surprise, diversion or intimacy of a system (Alben, 1996; Gaver and Martin, 2000; Norman, 2004) have gained increasing attention. But apart from becoming more and more focussed on the user and the interaction with the user, interfaces also have the tendency to become more and more integrated with each other. One can nowadays listen to music on a mobile phone and washing machines can dry the laundry (Thomas and Macredie, 2002).

Smart homes are a typical playground for the development of human computing applications (cf. Meyer and Rakotonirainy (2003)). Inside the house

users are observed by means of a large number of sensors and the pervasive, or ubiquitous, system looks after the house and its dwellers. Apart from adjusting the heating and light system, this smart environment could pro-actively close the curtains when it gets dark, alarm the police when intruders are spotted. One could even have a system that suggests which clothes to wear given the outside temperature (Intille et al., 2003).

Futuristic systems like these will be an excellent challenge and play-ground for researchers. There are, however, many aspects that need to be resolved before Human Computing, along with its ubiquitous interfaces, will really break through. Example hurdles include the precise and accurate recognition of events, the definition of optimal strategies to combine input from multiple sensors, and the development of performance metrics that can be used for system evaluation.

The development of natural interfaces that are able to perceive humans with their behavior through several modalities also comes at a cost. The perception of natural interaction requires systems to understand 'more' from human behavior, and besides the multi-modal aspects that are already mentioned, also the context-free and often ambiguous manner in which messages are created are to be taken into account. On the other hand, if interfaces become merged with everyday things, humans should be, or become, aware of the system, as they are initially ignorant (Nijholt et al., 2004).

Just these few aspects underline the difficulties associated with systems that are to provide useful responses to naturally communicating users. According to some scientists (Davies and Gellersens, 2002; Schmidt et al., 2005) many aspects of Human Computing even appear to be as futuristic today as they were in 1991. This thesis will, however, show that current state-of-the-art sensing technology is already very well able to comprehend humans and their communicative behavior at a variety of levels.

1.2 Meetings and Multi-Party Interaction

A domain for which the analysis and support of humans and their activities is relevant and practical is the domain of multi-party interaction and meetings. We cannot think of a world without them, and although sometimes we wish we could, they play an important part in our daily lives. Meetings are hard to avoid and everywhere. In the best case every meeting would be efficient, effective, manageable and with an outcome that is easily accessible afterwards. The reality though is that meetings are expensive, have an unpredictable outcome, prove hard to manage and usually hardly more than hastily written notes remain. Therefore it is not strange that research into the technological assistance of meetings and their quality dates over half a century back.

The domain embodies the comprehension of a subset of people's everyday activities, working and living, that moves beyond the individual. In multi-party interaction, messages are exchanged between individuals in various flavors and melodies, thereby exposing the full gamut of human communication abilities. Research in this domain represents a fundamental case, in which the automatic

analysis of human behavior provides value for social sciences and that opens the doors for the development of a wide variety of computational recognition techniques (Gatica-Perez, 2006).

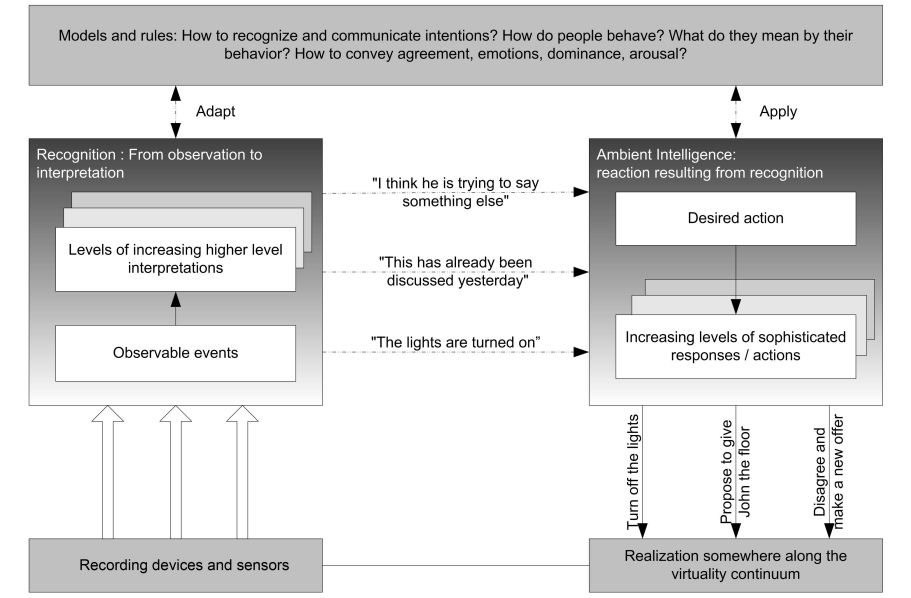


Figure 1.1: A schematic architecture of supportive technology from a human computing perspective.

Environments equipped with auxiliary devices such as microphones and data projectors have augmented 'traditional' face-to-face meetings for convenience as well as for the accessibility for interested others. The last decade, more and more smart meeting rooms appeared on the scene all of which are designed for information capture, information presentation and information interpretation. This is where human computing comes in. Every design of supportive technologies in the domain of multi-party interaction has one thing in common: it depends on interpreting incoming data, largely transmitted by humans, from a multitude of sensors. Where a simple system, e.g. one that is designed to switch off the lights, can be triggered by the start of a presentation, more complex systems will require information that cannot be obtained from direct observations. These more complex types of supportive systems need to reason about the information that is captured, possibly making use of contextual information, whilst applying predefined models of human behavior and fusing information channels to discover phenomena of interest. The aim of this thesis is to show to what extent it is possible to automatically obtain two of these so called, higher

level meeting phenomena (a meeting's influence hierarchy, and the meetings argument structures as they develop in a discussions) through the appliance of current state-of-the-art sensing technology.

Figure 1.1 shows the structure of a system that is designed to understand and also assist face-to-face communication. The examples in the model are tailored to the meeting domain and show various levels of complexity varying from signalling that lights are turned on, up to understanding that someone says something else than he, or she, actually means. The people and the environment are captured using cameras, microphones and other sensors. The resulting data are analyzed and recognition technology is used to detect directly observable events. This class of events embodies aspects such as speech, body postures, facial expressions and hand movements. Subsequent subsystems will analyze these observable events to fuse and transform them into progressively higher levels of interpretation (Reidsma et al., 2004). The raising of a hand, for instance, can be interpreted as a request for the floor. In a next, step the deduced interpretations are provided as input for the system's models that, depending on the abilities, provide suggestions for actions that can be performed. How the actual action is realized in the environment will be subject, amongst others, to the desired action type, the system's abilities and the environment.

1.3 The AMI Project

The research described in this thesis has been carried out within AMI[2], a European 6th Framework project. AMI is a 15-member multi-disciplinary consortium and short for Augmented Multi-Party Interaction. AMI targets the development of computer-enhanced technology that facilitates multi-modal interactions in the meetings domain. This and other projects on related subjects to multi-modal multi-party interaction have been started to bridge research on smart environments on the one hand and research on human-human interaction on the other hand. A multi-modal multi-party context is a context in which two or more persons interact with each other and/or with smart entities (objects, virtual humans, robots, etc.) present in that environment through a variety of information channels.

Within this context AMI aims to advance the state-of-the-art in areas such as human-human communication modelling, speech recognition, computer vision and multimedia indexing and retrieval. Its main aims are to develop technologies for the disclosure of meeting content and the provision of live meeting support. The provision of access to meeting data by means of a variety of tools that are able to retrieve relevant information for off-line and on-line browsing, including meeting structure analysis and summarizing functions fit without doubt, within the paradigms of interaction research, Human Computing, and Ambient Intelligence.

[2]http://www.amiproject.org

1.3.1 The AMI Meeting Corpus

One of the major deliverables of the AMI project is the development of a meeting corpus (Carletta et al., 2005; Carletta, 2006) aiming to benefit a range of research communities, including those working on speech, language, gesture, information retrieval, and object tracking, as well as organizational psychologists and sociologists interested in how groups of individuals work together as a team. Progress in these AMI research themes requires a large data set on which interaction research can be conducted, that allows for empirical observations and on which the foreseen technologies can be developed.

These requirements resulted in a scenario for design meetings in which four persons, as a project team, in a sequence of four meetings have to develop a design for a remote control (Post et al., 2004). Capturing meetings 'in the wild' would have resulted in a too diverged variety of meetings. The scenario was used to achieve controlled yet natural interaction between the meeting participants, rather than using predefined scripts that told participants explicitly what to do and how to behave. In the scenario, four participants play the roles of employees of an electronics company that has to develop a new type of television remote-control in order to create an attractive, user-friendly remote-control that could beat the unattractive and old-fashioned ones currently on the market. The participants were told that they were joining a design team whose task, over a day of individual work and group meetings, is to develop a prototype.

Design teams were chosen for the facts that: (1) The meetings had to be functional with clear goals, making it easier to measure effectiveness and efficiency. (2) Design is relevant for society. It is common and has clear economic value. (3) In design teams, the participants rely more heavily on information from previous meetings than in other types of teams. This dependency creates a good testing ground for investigating the possibilities of the browsing technology that is to be developed.

The participants were assigned four distinct roles: Project Manager (PM), Marketing Expert (ME), User Interface Designer (UI) and Industrial Designer (ID). See Figure 1.2 for a global camera view of such a meeting.

Over one hundred hours of meetings were recorded that followed the same scenario. The data were captured in meeting rooms equipped with many sensors, so called smart meeting rooms, at IDIAP (Martigny), at the University of Edinburgh, and at TNO Human Factors (Soesterberg).

Typical sensors that were used for capturing the data were cameras (recording global and close-up views), lapel microphones, microphone arrays, a whiteboard and smart pens. But also meta-information such as the seating arrangement, and the (powerpoint) presentations that were used have been collected. The recorded data, including layers of annotation (see Chapter 4) such as manually created transcripts, dialogue acts and summaries are all publicly available[3].

Before zooming in on the precise topic of this thesis, I will first provide a brief historical background of the area of technological meeting support by presenting an overview of previous and on-going related projects.

[3]http://corpus.amiproject.org

Figure 1.2: An overview image from one of the AMI meetings recorded at IDIAP

1.4 Related Efforts and Current Projects

Meetings, their technological support, and the process of human human interaction all have long been a subject of research (cf. (Licklider et al., 1968; Bales, 1950)). Looking at the subject from a historical perspective, it was Douglas Engelbart who foresaw the potential of the computer as a medium for idea development and group communication in the early 1960's (Engelbart, 1963b,a). Gains in productivity were predicted as a result of computational support. Not long thereafter, the notion that a computer actually could function as a medium able to dynamically transform information, rather than function as a repository, and to help people to share their view of the world with others, was presented by Licklider et al. (1968) along with the development of the communication network ARPANET. See e.g. (Treu, 1975) for an initial experiment that laid the foundations for an area nowadays known as Computer Supported Collaborative Work (CSCW). Also by the end of the 1960's and the beginning of the 1970's, Groups Support Systems in the area of decision support started to emerge (Scott Morton, 1971). CSCW has ever since focussed on increasing the effectiveness of work through the use of new media. In the mid 1970's a different, although

not completely distinct trend emerged. This trend focussed on the capabilities of technology to affect group interaction patterns from a sociological, but still task-oriented perspective (Vallee et al., 1975; Hiltz and Turoff, 1978). From the 1980's onwards, the social-emotional aspects of computer-mediated communication such as deindividuation, etiquette, and communication issues that arose due to lack of status and conversational cues when communicating by means of computers started to receive growing attention (Kiesler et al., 1984; Walther et al., 1994).

In 1987 it was Richman who predicted that software systems one day could change the way groups of people work together by means of *comprehending* the on-going group process (Richman, 1987). Although the state of the technology was far from actual recognition, the field of meeting analysis and augmentation by means of technology started to gain increasing momentum. In this same year two projects were launched that presented ideas to augment meetings with technologies that increased the participants' insights into the process, rather than just facilitating communication services such as terminals bulletin boards and email. One project was carried out by the MCC Technology corporation in Austin, Texas. This project became known as Project Nick (Cook et al., 1987) and concerned research into the development of meeting theories and the creation of meeting improving systems. The project enabled private, subgroup and public information transfer during meetings by making use of meeting rooms with individual displays and keyboards. Furthermore the display of *live meeting statistics*, such as a 'mood meter' is mentioned as well as the aim of creating a repository, or *public memory* of meetings. The other project, called CoLab (Stefik et al., 1987), was conducted at Xerox PARC in Palo Alto, California. Its focus was to make meetings more effective and to provide the opportunity for research on how computer tools affect the meeting process. The meeting tools that were devised to support the group interaction as well as the group's problem solving abilities were tools that allowed *parallel access to shared objects*. A tool called 'Cognoter' allowed for brainstorming, the collective preparation of a presentation, and for the organization of the meeting agenda. Cognoter was 'intended to *know-together*'. A second tool called 'Argnoter' facilitated the organization and evaluation of arguments for proposals.

By facilitating an increased insight into the process the role of technology as a static facilitator, did in essence not change. In publications on the CoLab project, the meaning of the word 'conversation' however, started to refer to "the combination of machines [..] and participants working together". In 2001, the NEEM project (Ellis et al., 2001, 2003; Barthelmess and Ellis, 2005) expanded the vision of conversing with machines in a way similar to Richman's predictions. The concept of autonomous software agents (See Section 3.3.2) was introduced into the meeting domain. These systems were to assist meetings on the informational, social and organizational dimension by means of adapting their actions to the environment dependently on their understanding of the environment. These sorts of systems, that adapt their actions to their interpretation of the sensed environmental information, all embody the domain of human computing.

All the projects mentioned spawned offspring at several institutions (see e.g. (Schultz et al., 2001; Garofolo et al., 2004; Morgan et al., 2003)), resulting in more and more technologies and recorded meeting corpora. In the last four to five years, there has been a real surge in the development. New large projects were established including consortia with partners from all over the world (M4[4], IM2[5], CHIL[6], AMI, CALO[7] and recently NECTAR[8]). All of these projects work on meeting collection, human meeting behavior, and on meeting supportive technologies. Projects consider design meetings like Project Nick, others focus on lectures and presentations. Some use natural meetings, others follow strict scripts or move somewhere in between. Recently even the question what people actually want from meetings has become a research topic in itself (Lisowska, 2003; Whittaker, 2005; Banerjee et al., 2005; Pallotta et al., 2006).

1.5 This Thesis

This thesis finds its origin in the AMI project and the assumption that the current state of technology in potential can overcome many drawbacks of the meetings of everyday life. Indeed, technology has had a significant impact on the way people can have a meeting. It has the ability to provide insights into a meeting, and it can even adapt itself to what is going on in a meeting.

The research in this thesis describes efforts in the direction of the automatic assessment of two higher level meeting phenomena in four person face-to-face meetings: influence hierarchies and argument structures. The assessment of higher-level, and more semantic, knowledge of a meeting is a broad research domain influenced on the one hand by the more sociologically oriented strains in group dynamics and conversation analysis (Bales et al., 1951; Sachs et al., 1974; McGrath, 1984), and on the other hand by the more computationally oriented approaches of concept and machine learning and their associated modelling techniques. The question if, and to which extent, these increasingly interwoven fields can aid us in the process to the automatic assessment of both higher-level group phenomena is investigated.

The actions that a prospective human-computing system might undertake, as a result of the actual recognition of the phenomena described, are chiefly subject to the (political and economical) operational environment. As a consequence, the impact of these actions, apart from some tentative explorations examining participants' responses, have been kept outside the scope of this thesis.

A general prerequisite for the automatic assessment of higher-level meeting phenomena is the existence of a model that does not just describe the phenomena at hand, but that also, when applied, provides sufficient structure for access.

[4] http://www.m4project.org
[5] http://www.im2.ch
[6] http://chil.server.de
[7] http://www.ai.sri.com/project/CALO
[8] http://www.nectar-research.net

I will show that, especially in the case of argument structures, the creation of such models alone, is by itself not a trivial task.

The automatic application of such a model, in turn, is to be carried out by algorithms that depend on the information signals that they have at their disposal. Ultimately, a system should be able to sense all the relevant signals, or features, required in a way that it unambiguously is able to apply the model. The issue here, however, is that one does not know beforehand which of the features that *can* be presented *should* be presented in order for an algorithm to be maximally successful. The quest for the ultimate combination of features for both phenomena is investigated by both a data driven approach (start digging into collected data for patterns and regularities), and a more sociologically inspired approach that formulates hypotheses and expectations based on existing literature.

Related work in the area of this thesis is for example reported in the areas of decision detection (Hsueh and Moore, 2007), action item detection (Purver et al., 2006) and group interest detection (Gatica-Perez et al., 2005).

The central question of this thesis considers *to what extent the current state of technology is able to automatically extract influence hierarchies and argument structures in four person face-to-face meetings.* Based on this main question the following sub questions are addressed as well:

- What are meetings, and why do they exist?
- How has technology influenced meetings up to the present day?
- What are opportunities and challenges for future meeting technology?
- How can we create systems that are able to comprehend more semantically oriented, or higher-level meeting phenomena?
- To what extent does current technology allow for automatic detection of an influence hierarchy from meeting participants?
- To what extent does current technology allow for the automatic acquisition of argument structures from meeting discussions?

The relevance of work in the area of technological support for group interaction and meetings in particular is rooted in both technology push and pull. The technology push coincides with the fact that Human Computing technologies, as described, have come to a point where a potential breakthrough in everyday meeting conduct can be realized. This, in combination with the fact that powerful and connected computers are appearing everywhere, more and more systems and applications are created that one should have in order to keep up and not to loose their competitive advantage. For meetings, potential applications exist in the areas of post-hoc retrieval, remote participation and real-time support. From an economic and technology-pull related perspective every development that increases (meeting) efficiency and effectiveness could be

worth the investment. Reduction of time and money with people jointly participating from all over the world, without the need to physically share the same location has already proven to be a viable market niche. A recent economic perspective even expects the market for rich media conferencing to grow 30% the next five years and to gain a total market value of over 6.5 billion US Dollars (WainHouse-Research, 2006). The recent attention of, and investments from governmental bodies, such as the European Commission and the U.S. Defence Advanced Research Projects Agency, that launched international projects signifies the potential importance as well as underlining the opportunities in this area.

1.6 Structure of this Thesis

The next chapter concerns meetings in everyday life and aims to provide a thorough and inspiring introduction in the main subject of this thesis. Although many people like meetings, there are numerous others who dislike them and cannot stop complaining. Some love to be amongst co-workers, where others doze off after the first few minutes. Why do people meet in the first place? How much time do people really spend in meetings?, and What do people want to know about them?

Chapter 3 elaborates quite generally on the opportunities technology has created for improvements in the meeting domain and the consequences this has involved. The main aim is to sketch possible application domains for human-computing technologies, especially those described in the subsequent chapters. Chapter 3 describes ongoing developments in pre-meeting technologies, such as meeting scheduling systems that assist in the organization of the actual event while adhering to the wishes and constraints of the *sensed* environment. Also, the opportunities for, and current state of the art in technology for real-time meeting assistance is charted. A specific focus lies on the need, and possibilities, for adaptive systems that are able to induce higher level meeting phenomena. Examples in the areas of pro-active agents and group support systems are discussed. The last area of technological support that is described can be used once a meeting is over. Meeting browser systems that can provide (automatically) derived meeting information are discussed. The chapter will conclude with a section on how technology has influenced the behavior of the participants themselves. It is investigated how these changes have influenced the actual meeting process and meeting outcome in such a way that opportunities for future applications become clear.

For all these meeting supportive technologies the question is how to create and apply models that can equip systems with sufficient knowledge of the environment to fulfill their projected tasks. Chapter 4 describes the applied methodology used in subsequent chapters to model, and gain automatic access to, the concepts of influence and argumentation. The method in essence uses knowledge from existing literature in combination with a corpus of signal recordings, to derive detectable aspects that are related to the concepts of interest.

Initially, a model, or annotation schema that aims to capture the concept is manually applied on the corpus. The resulting examples, or class labels, that describe the phenomena of interest are then combined with the set of possibly related aspects, or features. Machine learning algorithms are then released onto this data with the aim to replicate the manually defined class labels, as defined by the annotation schema. Finally, in a optional step, the set of features that has been used to predict the class labels is then reduced to reach an optimal number with respect to the replication error and obtainment investment.

The methodology described is adopted to assess influence rankings of meeting participants in Chapter 5. First, related theories such as Social Status Theory (Berger et al., 1980) are introduced to provide an introduction to the concepts of influence and dominance and this way shed light on potentially relevant features that can aid the classification process. Two attempts in the direction of automatic hierarchy replication are described. In the first attempt, class labels were created from observation by people who did not participate in the examined meetings. For the second experiment, the labels were deduced from questionnaires issued to the meeting participants themselves. The feature set was expanded in the second round as more features were available by that time and the results of several types of classifiers were compared. The resulting system capabilities are transferred into a prototype meeting browser and a 3D meeting environment.

A second adoption of the methodology is presented in Chapter 6, with the aim to end up with a system that is able to automatically create argument diagrams of meeting discussions. The chapter starts, in line with the previous chapter, with an elaborate description of several descriptive theories before the development of our own annotation schema is described. This schema, that attempts to structure a discussion, can, apart from functioning as organizational memory, also be used as interface in a meeting browser or function by itself as a feature for other algorithms, such as those that try to summarize a meeting. Two important steps towards automatic application of this schema are investigated: the automatic labelling of the various speaker contributions within a discussion, and the labelling of pre-identified relations between these contributions. A user experiment is reported that investigates the useability of the diagramming method in the context of answering questions related to the debates. Finally a prototype application is presented into a meeting browser that allows users to navigate through the debates.

As one could expect influence and argumentation to be somehow related, an exploration of mutual dependency is carried out in in Chapter 7. People with various levels of influence are examined during several discussions and possible differences are explored with respect to the frequencies of the categories defined in the argument structure. Furthermore aspects such as differences in turn duration, as well as differences in the number of contributions, are investigated. The chapter also contains a section on rule induction, an unsupervised approach to reveal interesting associations and correlation relationships between aspects of both phenomena. Then, the results on the classification performances for both phenomena are explored after using one phenomenon as a feature of the

other and vice versa. The chapter finishes with the creation of a profile of the behavior of influential participants in meeting discussions.

Having investigated in which way and to what extent the higher-level phenomena of influence hierarchies and argumentation structures can be assessed, Chapter 8 steps down and zooms out of the subject back again to the level of technology-aided meeting assistance. This time the focus is not to report on ongoing developments but rather to look beyond the current state of technology into the future of meeting assistance. The chapter elaborates on the emerging developments in 3D remote physical appearance. An experiment is described that accompanied meetings with a virtual meeting chairman, that steered by a wizard imitated behavior that potentially can result from the models created in the previous chapters. The chapter finishes with a section on ethical implications and considerations, and a section on the challenges that need to be resolved before the emergent meeting technologies described can be maximally exploited.

In the concluding Chapter 9, the answers to the challenges and research questions are given. The chapter touches on the progress that was made in the area in general and on the insights that were gained during the research for and writing of this thesis.

Chapter 2

Meetings of Everyday Life

A meeting is a place where you keep the minutes and throw away the hours. (Thomas Kayser, 1990)

2.1 Introduction

Meetings come in all sorts and flavors. A meeting can be successful, boring, scheduled or organized. Meetings can be a platform for groups to interact, to exchange thoughts, to collaborate and to move things forward. They belong to the way we organize our work and in fact, the world would be very different without them. The general phenomenon 'meeting' can be described as an organized group process where people collectively engage in an activity of communicating information in order to serve a common goal, for example to make decisions, to resolve a dispute, or to come up with a new product. A meeting is a realization of what (Clark, 1996) called a joint activity; it involves two or more participants that interact.

At first glance, one would say that meetings are one of the most well-understood phenomena in society. They are everywhere, so they are likely to be well understood. That assumption however, does not seem to be true at all. One reason for this is mentioned by Schwartzman (1989), who states that exactly due to their pervasiveness and the fact that they are taken-for-granted in everyday life and within organizations, they have not gained much attention. From a research perspective, meetings have mainly functioned as a testing ground for theoretical models in the field of small group research [1]. In an overview of 30 years of research interests within this field, Zander (1979) mentions the cohesiveness of groups, the nature of social pressure, and the dynamics of making group decisions as the most dominant topics. From an organizational perspective meetings have primarily been regarded as a management tool. A tool that, like any other, requires optimal usage to be maximally effective (Doyle

[1] See Bales (1950) for an example

and Straus, 1976; Burleson, 1990; West, 2003) with much effort spent in investigating the effects of power and leadership issues (Sell et al., 2004; Paulsen, 2004).

This chapter will discuss the meetings of everyday life with a focus on business meetings. The aim of this chapter in the context of this thesis is to provide a general background for the main topic of this thesis: face-to-face meetings. Section 2.2 addresses the intrinsic human drive to form groups and discusses some benefits and potential pitfalls that are inherent to meetings and collaboration in general. It explores facets of humans engaged in interaction rituals that result from communication protocols pertaining to the frame of human-human interaction. Section 2.3 then goes more into depth on the meetings central to this thesis, the business meetings. The important aspects of everyday meetings are charted in terms of input, process and output, and profiles of typical business meetings are given. Do people really lose interest, and what are the resulting consequences in terms of the meeting outcome? Section 2.4 explains when meetings are successful and when they are not. Attention is given to meeting behavior and the 'rules' of conduct one should address when a meeting is to be successful.

In essence this chapter provides the ground for the next chapter, as it will highlight most aspects and problems associated with everyday business meetings and as a result points out possibilities for technology to enhance and complement meetings and the way meetings are perceived.

2.2 Why people work together

Children in our current day society are taught to play together, in such a way that they develop skills for later life so that they can live and work with others. Living and working with others, or the human ability to form groups naturally have been mentioned as a characteristic of the human being (See (Coon, 1946)). The need to be part of a group is innate to humans and part of their biological inheritance (Baumeister and Leary, 1995) and the formation of small groups have proven to be the basic survival strategy for the human species. External threats, such as a shortage of food or defence against rival groups increased group cohesion and gave individuals a competitive advantage. Nowadays encounters between groups generally still have a confrontational nature (Hoyle et al., 1989). It is undoubtedly true that the ability to work together has resulted in an astonishing progress of the human kind and facilitated a more efficient and effective execution of many human activities (West, 2003; Weldon and Weingart, 1993). Goals have been accomplished by means of resource pooling and risks and costs have been shared as labor, knowledge, abilities, experience, time and money has been combined.

From an organizational viewpoint, along with the division of labor and the globalization of markets, the work within companies and organizations has grown more and more complex and requires an increasing need for coordination and structure. This, combined with the fact that product development

times are shortened due to market pressure, has led organizations to move away from hierarchical forms into more organic and flattened forms. As a result of this teams have become more and more the building blocks of modern organizations and increasingly started to become everyday practice (West, 2003; March and Sevon, 1984; Appelbaum, 1994). Preferably people with different backgrounds and unequal status are put together, aiming for the cross-fertilization of ideas, which in turn should result in high quality decision making, creativity and innovation. (West, 2002; West et al., 2003)

2.2.1 The benefits of working together

Nunamaker et al. (1991) identified five major gains for working together in a team or a group. In the first place, as skills and knowledge are pooled together, a group as a whole has more information than any individual member by itself. Second, when information is exchanged, team members might use this information in different ways, reasoning from a variety of experiences and backgrounds. Third, due to the presence of others, individual errors are noticed more easily. Fourth, the group members improve their performance by learning from and imitating the more skilled members. Finally, the sense of being part of a group may encourage and stimulate individuals to perform better as there is an increased opportunity for recognition by others (Hellriegel et al., 1995), and generally more responsibility associated to the task (West, 2003). A nice example where the benefits of group work are apparent is described by Slavin (1983). He showed that if students work in groups, rather than individually, they work harder, help less-able group members, and learn more.

Reading the above, one would expect that working in groups gives results that are exceeding the sum of its individual members' contributions. The director of a company where I conducted my internship once quoted his former boss evaluating him and his associate: "You, and You make Eleven.", he said whilst pointing subsequently at my boss and his associate before combining the two pointing fingers into the figure of eleven. This however is not necessarily the case. In fact there are many barriers to overcome before teamwork is more effective than the work of the individuals combined.

2.2.2 The drawbacks of working together

A famous example that shows drawbacks of team work are the Ringelmann experiments (Kravitz and Martin, 1986), where students were instructed to pull a rope as hard as they could. The force on the rope was measured. In a second round teams of students were to pull that same rope. It turned out that the teams were pulling around 75% as hard as the aggregated work of the individuals.

The phenomenon where individuals hide themselves behind others, and thereby put in less effort, is known as social loafing (Latan et al., 1979). Social loafing can be a result of de-individualization, the fact that people have problems with making personal goals subordinate to group goals, or the need to compete for

airtime (Hellriegel et al., 1995). Nunamaker et al. (1991) lists further pitfalls related to the unavoidable division of time amongst participants. The absorption and remembrance of ideas from others limits the time members have to think for themselves. Members may lack focus and miss or forget contributions, or information may be presented faster than that it can be processed. Inappropriate communication strategies, or meeting domination by some group members might even prevent members from contributing, resulting in a potential loss of ideas (Hirokawa and Pace, 1983; Hellriegel et al., 1995). Although necessary for effective functioning, the requirement of non-task discussions, or meta-level communication, (e.g. related to the communication strategy) is time consuming and thereby reducing the performance.

Another threat for the group to function are social issues related to what Goffman (1955) described as preservation of *face*, or the image that members of themselves try to preserve in relation to others. The potential loss of face as a result of a negative evaluation can cause members to withhold ideas and comments, and to be reluctant to criticize the comments of others due to politeness or fear of reprisals. When members refrain from deviating contributions, the threat emerges that discussions follow just one single train of thought.

2.3 Meeting aspects and meeting behavior

Having a meeting, or in the terminology of Clark (1996) having a joint activity, is where the group accepts expectations and obligations and where the participants interact with one another. When interaction takes place, the main medium of expression is talk (including non-verbal talk). Habermas (1984)'s theory of communicative actions for instance states that for an (communicative) action to take place, all persons involved should have the possibility to apply speech acts. The ideal speech situation, he states, is oriented towards rational argumentation, where participants have equal rights to state, to question, to request and to criticize an there should be an symmetrical distribution of opportunity for all participants to choose and to practice speech acts. Preferably this process should not be impeded by personal preferences, emotions, power, etc.

It is obvious that in meetings and everyday conversations this cannot be achieved simultaneously. It is therefore the aim to jointly achieve a communicative action confronted with the constraints of everyday life. Bales et al. (1951) for instance, observed that as groups work together, social and emotional differences such as conflicting values, cultural norms, and different methods of expression emerge. These conflicts can hinder the group and, as a consequence, behavior to reduce the tension appears. It is the frame of the interaction, as Goffman (1974) has put it, that specifies the norms of the interaction in which people follow agreed patterns of conduct. Or the other way around, as stated by Orlikowski and Yates (1994): The modes of acceptable conduct form an established repertoire that exist for each form of communication.

2.3.1 Meeting behavior

Over the years people have developed all sorts of norms and tools to regulate the meeting process. Where in the early days, when norms and rules were still in a developmental phase, a two-bladed fighting axe was used in folk meetings to execute law breakers, or to chop off limbs or pieces of the clothing of those who violated meeting rules (Van Vree, 1999), nowadays the sharp axe has developed into a small gavel and even this is hardly used anymore. Another example is the fact that people generally sit when they meet. A reason for this could be that it is more convenient to sit than to stand. From an evolutionary perspective it is rather awkward, as stand-up meetings appear to be faster and produce decisions of similar quality in comparison to sit-down meetings (Bluedorn et al., 1999). The seating arrangement itself, however, has started to play an important part in the meeting process. A common assumption is that a round table facilitates discussions and balances hierarchy, whereas a rectangular table emphasizes the hierarchy and leadership (Burleson, 1990). Other mechanisms for control are, apart from leadership issues, the usage of an agenda, and also the formation of coalitions to eliminate divergency. Alvehus (1999) described the problem of overcoming the recurrent problems in group interaction as the process of turning meetings into machines that are expected to efficiently transform input into output.

The displayed behavior of meeting participants can be steered by tools, but the evaluation of the behavior itself is relative to social norms that are generally unstated and unwritten. Typical forms of social norms one might encounter are that yelling and screaming are unappreciated, one should let people finish talking, no private conversations, no whispering, and one should refrain from 'Ad Hominem' arguments. Also emotionality must be reduced to have a meeting more suited for problem solving as ideas should be separated from the person and opinions are said to be less interesting than facts (Burleson, 1990). Grice (1975) formulated in this respect four maxims that hold for cooperative conversations. The maxims of quantity, quality, relevance and manner state that one should say nothing more or less than is required, one should speak the truth, one should only say things for which one has enough evidence and that are relevant for the discussion at hand, and finally one should formulate such, that it can be easily heard and understood by the interlocutors.

Meetings are nowadays often facilitated by a chairman, that embody the tool for regulation. A meeting chair is guided by the developed norms and a typical chair should e.g. facilitate the participants to have their say, cut off people who make their contribution too long and intervene when contributions are not relevant to the discussion at hand. Also, discussions should be properly organized to have arguments develop, so that all positions are put to the fore, and all relevant pros and cons are raised. All these norms and conventions define the shared belief of what is normal and acceptable and hence constrain people's actions aiming and steering for a successful outcome. To act in line with the social rules and norms (face goal) is often conflicting with the wish of a participant to immediately achieve their agenda or objective (task goal). For the meeting

process to function, a balance between these two levels of communication has to be maintained (Tracy and Coupland, 1990).

The process of conversation is thus, to a large extent, rule-bound, and the participants are highly skilled at respecting and adhering to the rules of the game. In terms of regulating talk in conversations, for example, the process of turn-taking, or the way participants manage turn-change, is a typical system of practices, conventions and procedural rules that function as a means of guiding and organizing the conversational flow (Goffman, 1955). With the basic rule for conversation being one party at a time (Schegloff, 1968), the process of turn-taking is a typical mechanism that prevents people from bumping into each other during conversations (Duncan, 1972). Speaker turns, interruptions, and passing the floor from one speaker to the next are accomplished in a variety of subtle and mutually understood ways. Duncan (1972); Duncan and Niederehe (1974), for instance, studied how different cues signal the intention of participants to either keep, take or yield the turn. Different interpersonal affiliations result in differences in conversational sequences, but also non verbal aspects, such as gaze, facial expressions, posture, head movements, gestures, and many others contribute in their own way to the flow of conversation and the human perception of social interaction (Argyle et al., 1973). As we will see over the next chapters, a thorough understanding of these issues is required when aiming to assist this process by means of technology.

2.3.2 Framing the concept

Meetings range from scheduled, formal meetings of a corporate body (board of directors meeting) to informal, ad hoc two person meetings of colleagues that address a specific issue by the coffee machine. The previous paragraph showed a number of aspects that inherently relate to the concept of meetings and meeting performance and that explain the huge variety of the everyday meetings that we encounter. Amongst these, the participants that constitute the group of the meeting, the social norms and regulations the group takes into account and the importance of the setting have been mentioned. To provide a more comprehendible and complete picture Dennis et al. (1988) created an overview of meeting related aspects that can influence the performance. An interpretation of this overview, or model, is shown in Figure 2.1.

The aspects mentioned can be taken into consideration when one, for example, wants to compare two meetings.

Central to the model is the meeting process. The process creates the output given a group, a task, and a context. The meeting process structures the activities that the participants execute to achieve the goal(s) related to the tasks at hand. The form and the degree of structure that is applied, the equality of participation, and the level of conflict are a part of this. Antunes and Carrio (2003) describe three possible approaches for creating a typology of meeting processes: the genre approach, the decomposition approach and the individual intervention approach. The genre approach focusses on the purposes and the communication patterns based on recurrent communicative actions such as a

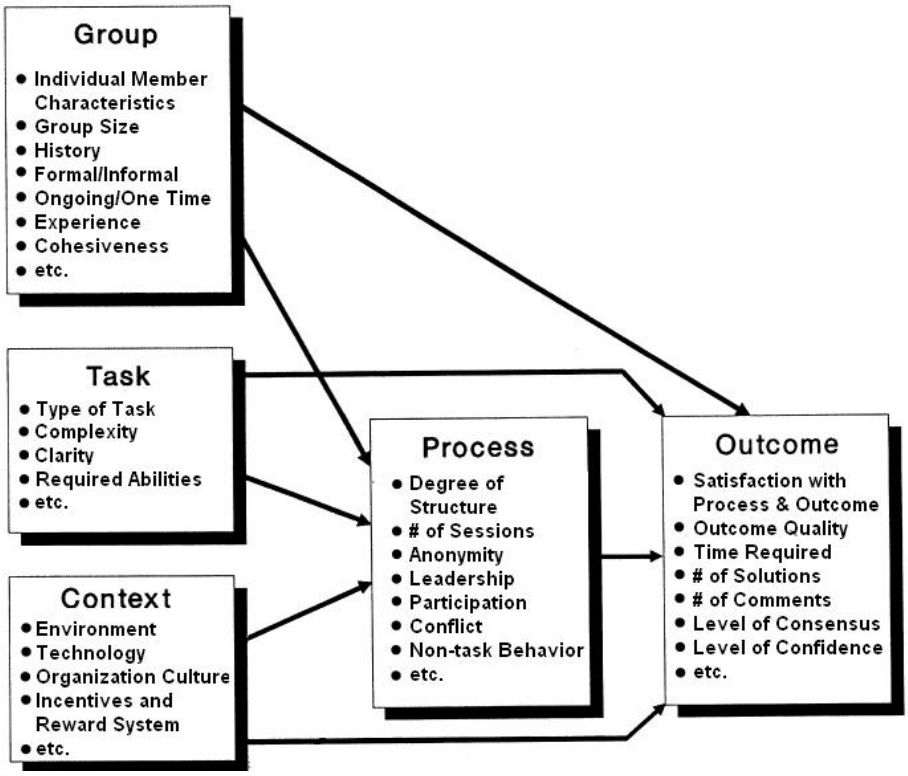


Figure 2.1: Important meeting aspects. (Inspired by Dennis et al. (1988))

briefing, a progress report meeting and a brainstorm. The decomposition approach considers meetings as decomposable into multiple levels of detail with goals and sub goals, such as a recursive combination of divergent, convergent and closure phases. The individual intervention approach structures the meeting process according to individual (process or task related) interventions produced by the participants. Examples are: defining the agenda, opening and closing the meeting and making a statement. A whole different kind of typology is offered by McGrath (1984). He defines a process to be either a process of generation (making a planning, or being creative), a process of choosing (intellectual, decision making), a process of negotiation (resolving conflict), or a process of execution (a performance, contest, or battle).

With respect to the group one can consider the characteristics of the individual members and the related experiences such as skills and abilities. Also the motivational factors, the cohesiveness and the size of the group play a part. A possible classification could distinguish between *Primary groups* and *Secondary groups*. Primary groups consist of small groups with intimate, kin-based relationships: families, for example. They commonly last for years and are small.

Secondary groups are the foremost large groups where relationships are formal and institutional. People here are brought together to perform specific tasks of a non-routine nature before being disbanded (Cohen, 1993). The formation of primary groups, of course, can very well happen within secondary groups. A different typology could focus instead on the group's ideologies with possible categories such as conservative, moderate and liberal.

According to Hoffmann (1979), there are three types of individual behavioral roles that can be identified in groups or teams. These roles can be classified as task-oriented, relation-oriented and self-oriented. Each group member has the potential of performing all of these roles over time. *Initiators, Coordinators* and *Information Givers* are task-oriented roles that facilitate and coordinate the decision making tasks. The Relations-Oriented role of members deals with team-centered tasks, sentiments and viewpoints. Typical examples are : *Harmonizers, Gatekeepers* and *Followers*. The Self-Oriented role of members focusses on the members' individual needs, possibly at expense of the team or group. Examples here are *Blockers, Recognition Seekers* and *Dominators*. The Dominator is a group member trying to assert authority by manipulating the group or certain individuals in the group. Dominators may use flattery or proclaim their superior status to gain attention and interrupt contributions of others. According to Hellriegel et al. (1995), a group dominated by individuals who are performing self-oriented sub-roles is likely to be ineffective.

Task characteristics largely determine the amount and type of information that will be exchanged in a meeting and it is therefore not strange that tasks are said to account for 50% of the variance in group performances (Poole and McPhee, 1985). When zooming in on the characteristics of a task, the task complexity, the task adaptability (how transferrable is the task?) the usability of the task (how easily can the task be learned?) and the task clarity come into play. But a task usually also has constraints, such as time constraints, and goals that can be quantitative, or qualitative. A task can require abilities of the performers, such as knowledge, (behavioral) skills, and materials. Jonassen (2000) described eleven types of tasks. The task types vary from logical tasks that are known to have a clear specific solution (e.g. solving the Towers of Hanoi problem) and an algorithmic task (e.g. solving a math question) all the way down to design tasks that have an unclear outcome with ambiguous solutions (e.g. design a kitchen), and dilemmas where no correct solution might exist(e.g. finding an answer to the question whether euthanasia should be legalized?). Steiner (1972), by contrast, described tasks by comparing the productivity of groups in relation to that of individuals. In additive tasks the contributions of each member are combined into the final group product. Lifting a couch, for example, involves the efforts of everyone, and the group is more effective than an individual. In disjunctive tasks, a single person can find the solution. A group of people working on a crossword puzzle is, although more likely to solve it as a collective, dependent on the contributions of the individuals to specific solutions. The third type Steiner (1972) identifies is the conjunctive task, for this type of task the productivity is limited to the least competent member. If a group is tied together when climbing a mountain the performance depends on

the weakest link in the chain.

It should be noted that interdependencies exist between the task and the group. Aspects such as familiarity, salience, and the belief in success can have a direct impact on the motivation, and on the performance.

The context relates to the resources of the meeting, or the conditions and factors that are given. Here the environment plays a big role, including the generic facilities such as the location, the table setting and the physical working conditions. According to Drew (1994), the place where people meet is as crucial as why and when people meet. Also the information resources that are available to a group to accomplish the task have to be considered, such as the existence of an agenda and other supporting documents. Not to forget the existence of formal organizational rules, influences of higher management, and the existence of the possible rewards and punishments. The context thereby determines possible risks and matters of urgency involved. The most important part of the meeting context however, and I could not any longer avoid mentioning, is the presence or absence of supporting technologies. The growing array of technologies nowadays fundamentally shape the ways that meetings take place (Hellriegel et al., 1995) as technology offers new perspectives on, amongst other things, communication and language, human perception and social interaction. Technology that aims to aid the meetings is sometimes referred to as Groupware (Grudin, 1994; Leventhal, 1995; Nunamaker Jr. et al., 1995), CSCW (Computer Supported Collaborative Work) (Grudin, 1988; Monk et al., 1996) or GSS (Group Support Systems) (Briggs and Vreede, 2001; De Vreede et al., 2003) and comprises both communication tools, such as electronic brainstorming, as well as specially designed physical facilities, such as large displays (See Chapter 3).

The final meeting aspect we address is the meeting outcome. The outcome of a meeting involves the change of group, the task and the environment in which the meeting took place. All of these are a result of the process that in turn can also be evaluated *exempli gratia* as 'interesting', or 'boring'. The transformation of a group occurs when a group, for instance, becomes more or less cohesive or the hierarchy within the group changes. Changes in the environment can be a result of a meeting as chairs can be rearranged, or rooms redecorated.

It is important here to consider that the model addressed is certainly not the perfect model as the perfect model does probably not exist. For an alternative, similar model, one can consider, for example (Pinsonneault and Kraemer, 1989).

Generally, however, a meeting is evaluated with respect to the completion of the task. This can be assessed both qualitatively and quantitatively and a meeting, for example, could turn out to have been a 'good' meeting when ten out of twelve solutions emerged. People can be interested in efficiency measures, such as the number of alternatives that was discussed regarding particular solutions, or in the average time that were required for a decision to be made. Others could show interest in measures that relate to the effectiveness of the meeting and they, for instance, could be willing to assess the quality of the discussions, the quality of the generated solutions, the quality of the decisions that were made or the resulting attitudes of the participants. These attitudes contain subjective information regarding all the major meeting aspects including the

group's functioning, the process, the context, and the task.

2.3.3 Meeting Profiles

The profile of a typical meeting in corporate America is a staff meeting, held in a conference room, starting at 11:00 am, lasting one hour and thirty minutes, involving nine people, with no written agenda, an atmosphere between somewhat to very informal, and is a meeting where eleven percent of the time is spent discussing irrelevant issues (Mongue et al., 1989). We will go into more depth here about findings related to the considered meeting aspects of the previous paragraph.

A trend that can be observed over the years is that there generally is an increase in the number of meetings employers have. Along with this, the time that people spend in meetings has increased as well (Romano Jr. and Nunamaker Jr., 2001). Mosvick and Nelson (1987) even reports that average executives in the 1980's participated nearly twice as often in meetings as in the 1960's.

With respect to group size, it was found by Slater (1958) that members of a group of six or smaller never felt their group too large and that members of a group of four or larger never felt their group too small. As a result he predicted five to be the optimum group size when exposed to an intellectual task "representing the most common variety faced by groups in everyday life". By the mid 1980's Mosvick and Nelson (1987) confirmed that the ideal size of a meeting is either five or seven people. Groups smaller than five lack the expertise to handle tasks efficiently, whereas groups larger than seven start to have increased problems with controlling the group's dynamics. The notion that the larger a meeting is, the more structure it requires is backed by Doyle and Straus (1976). However, they also note that the optimal meeting size is dependent on the purpose or task that is to be conducted. (Drew, 1994) lists optimal group sizes for seven different meeting types. For problem solving and decision making the recommendation is to have five or fewer participants. Problem identification meetings should be held with ten people, and a training seminar with around fifteen. Informational meetings, reviews, and presentations can be held with up to thirty people and for motivational meetings they state that the more participants there are, the better.

Although actual numbers are scarce, Panko and Kinney (1995) report that when considering 446 oral communication episodes of 53 MBA students with a full time managerial position, dyads are by far the most frequent and comprise a share of 65%. Dyads and triads together accounted for 75% of all meetings. Mongue et al. (1989) reports in a survey over 900, three or more person, corporate oriented, meetings that 20% had fewer than six participants, 41% were in the range from six to ten and 22% had sixteen or more participants. Although dyads were not taken into account here, the reported figures appear to be rather indicative as they do not seem to align with those mentioned by Panko and Kinney (1995).

With respect to the general meeting task or the purpose of coming together Mongue et al. (1989) report that most of them were held to resolve conflicts

(26%) or to reach a group decision (26%). Other reported tasks include solving a problem (11%), ensuring that everyone understands (13%), gaining support for a program or report (7%), the facilitation of staff communication (5%) and the exploration of new ideas and concepts (4%). From these meetings 52% were staff meetings, 22% task force meetings, 21% information sharing meetings and 5% brainstorming meetings. The remaining 7% were classified as 'other' meetings.

When considering the meeting environment, within business organizations the most frequently reported sites are conference rooms, offices, hallways, restaurants and cafeterias and breakout rooms. Panko and Kinney (1995) report that most of their meetings took place in offices (50%) and conference rooms (26%). Most of the meeting time however was spent in conference rooms (54%) rather than in offices (28%). From the time that was spent in dyads they report that by far most of the time (74%) is spent in offices. When omitting dyads, the time in conference rooms increases to 67% and drops for offices to 16%. This is more or less in line with the figures reported by Mongue et al. (1989), who report 74% of their examined corporate meetings to have taken place in conference rooms, and 15% in offices.

Regarding the duration, Panko and Kinney (1995) report that out of all their meetings 75% did not last longer than thirty minutes, 28% was even shorter than five minutes and 3% lasted longer than two hours. The percentage that is taken by meetings lasting two or more hours, constitutes as much as 50% of all the meeting time, whereas this is just 3% for meetings shorter than five minutes. For the corporate meetings of three or more people described by Mongue et al. (1989) the most frequent duration was one hour and thirty minutes, 27% of the meetings lasted shorter than one hour, 41% between one and two hours and 10% of all meetings lasted over four hours.

Having seen these figures, they hardly say anything about the achieved level of success or failure of an average meeting. The next section will address this, and focusses on more influential factors in relation to meeting evaluation.

2.4 Meetings: a Love-Hate Relationship

When one uses the term 'meeting' in an ordinary conversation, there are chances of responses that express a certain degree of disdain. Indeed, unsuccessful meetings are not known to be an exception (Romano Jr. and Nunamaker Jr., 2001). In this section, we will highlight aspects that, more than once, have been mentioned to play a part in the successful outcome of a meeting and the associated level of satisfaction achieved by the participants before stressing the need for ways of assistance that overcome the negative aspects associated.

A good meeting can be defined as one in which both organizational and personal goals are achieved and the social well-being of a group and its participants is maintained. Meeting failure on the other hand is suggested when the results and effects of a meeting are diverse and unfocused, ranging from inadequate meeting minutes, vague action items, feelings of wasted time and disrupted bonds amongst team members. All these effects can be very difficult

to capture and quantify. One could try to evaluate in terms of the outcome and regard the extent to which the generated substance, the actions, solutions and decisions met the objectives, or address the way the problems and potential solutions were identified. Another way is to evaluate in terms of the process and focus, for instance, on the extent to which the group acted as a team. Wynn (1979) mentioned the chances for equal participation and the willingness to share information and ideas as factors for success.

Generally there are two types of evaluative aspects. The outcome and the process. This relates to a finding from Bales (1950), known as the equilibrium theory, stating that work on the group is as important as work on the task, not to forget the work on the context.

2.4.1 When meetings are successful

The chances for the success of a meeting highly differ per meeting and are highly related to the level of preparation. Some important points are discussed here.

In the first place, one should thoroughly consider a meeting's legitimacy beforehand and investigate if the projected 'costs' will justify the expectations. Alternatives should, for example, be considered if not all relevant participants, those who can make decisions), can be present and if not all relevant information is available.

Secondly, control theory (Carver and Scheier, 1990) suggests that if the perceived performance meets or outclasses the expectations, positive feelings ensue. In other words, people will like meetings, when the process is good and the prospected results are obtained (O'Connell et al., 1990). So, where the provision of information before a meeting takes place, these expectations can be steered. The appropriate provision of information (covering as many aspects of the meeting as possible, including the group, the task, the environment, the process and the projected outcome) will, as a result, increase the chances for success (cf. Robert (2000); Hocking (1996)). This pre-meeting preparation identifies the frame of the interaction in an early stage and as a consequence rules out surprises, constrains the possibilities, and provides a focus. The predefinition of the meeting objectives and time-frame, for example, give people an idea of the expected efforts. Information about the location and agenda are clues for the meeting atmosphere and provide the opportunity to become familiar with the topics that are to be discussed.

Another possible factor of success is the presence of leadership. Leadership in meetings is commonly institutionalized by the appointment of a meeting chairman who is responsible for the preparation phase and expected to guard the task and the group along the meeting process. The chairman can influence the task (e.g. predefine the agenda), the environment (choose a suitable location where, for example, external noise is reduced) and the process (begin and end on time, assure a balanced participation and make sure the agenda is followed). The success of the meeting, however, is not fully dependent on a chairman's presence. (There could even be none.) The participants are responsible themselves, they are expected to appropriately participate and to display good meeting conduct

within the constraints imposed by the environment; possibly including those from the chairman, (see also Section 2.3.1). Examples listed in various sources mention, amongst other things, the willingness to cooperate, the willingness to share ideas, the willingness to stick to one plenary conversation, the willingness to focus all comments on agenda items, the willingness to refrain from interruptions and personal attacks and the obedience of participants' roles such as the meeting chair (Van Vree, 1999; Robert, 2000; Hocking, 1996; Drew, 1994; Doyle and Straus, 1976; Burleson, 1990). The level of synergy achieved between the leader and the group can, as a result, be another factor to success.

Also, a variety of techniques can be employed for various activities that can take place in a meeting that have proven to improve the outcome. Examples of these activities are monologues, discussions, presentations and brainstorm sessions. A typical brainstorming technique to make sure everyone is able to contribute is the silent writing down of ideas (cf. Doyle and Straus (1976); Burleson (1990)). Typical presentation techniques are to speak up and to face the audience at all times.

Once a meeting is over, for the participants it is important that the contents, or *group memory* created during the meeting, including solutions, decisions and action items, are prevented from disappearing and becoming accessible to the public (see also Section 3.4.2). This step completes the meeting process and discloses the content by providing accessibility to whomever it may concern.

2.4.2 When meetings are unsuccessful

The task of organizing and executing an effective meeting can, however, be both time consuming and difficult. Even after extensive preparation, there are no guarantees that a meeting will proceed smoothly nor that it will reach the desired goals. Reasons can be related to the unpredictable behavior of participants, the lack of structure, and the lack of preparation. Rogelberg et al. (2006) even found that individualistic oriented employees conceive meetings as interruptions which, by nature, have a negative impact on the well-being of employees.

In principle one could say that the aspects identified above that can contribute to the success of a meeting, can also be the source of meeting failure. Lack of notification, lack of individual preparation, lack of an agenda and lack of control can all be sources of unsuccessful meetings. If a meeting is not necessary people will feel as though their time is being wasted. They will refrain from active participation or seek refuge in different subjects. If the purpose, or the task, is unclear, participants cannot prepare properly. If the wrong people are at the meeting, the input from the people will be of less value and a wrong setting, finally, might disrupt the process from taking place at all.

With respect to the group, size can be a factor. Dependent on the meeting genre, the meeting group size is important in order to maintain order and make sure everybody is able to have his or her say. Related aspects in this sense are leadership on the one hand and monopolization of dominant participants on the other (Hellriegel et al., 1995). Due to time constraints a meeting can be under

high pressure and people leap into problem solving before it is clear what the actual problem is (Doyle and Straus, 1976). Confusion might arise about which roles there are during a meeting and which responsibilities are associated with them.

The most important lesson is that meetings are often inefficient (Romano Jr. and Nunamaker Jr., 2001). Gordon (1985) found that in up to 50% of the meetings productivity is wasted. A large study of business meetings in the UK (MCI WorldCom, 1998) shows that 80% of the professionals, who meet on a regular basis, admit to daydreaming, 23% admit to have dozed off, and almost all of them have missed meetings. Furthermore, when looking at meeting resources, for example, they are notably expensive. According to MCI WorldCom (1998) a typical busy professional attends nearly 60 meetings a month, of which more than 10% involve travel out of town. A typical out-of-town six-person meeting costs £1.645, including significant soft costs, such as the loss in productivity during travel and while arranging meetings. All in all, business organizations are said to spend on average around 7 to 15% of their budget on meetings directly.

The general inefficiency of meetings, the fact that they are expensive and the fact that people have to travel to reach a meeting, combined with the fact that they are pervasive and we cannot do without them, lay the foundations for efforts into meeting improvement. The next chapter will assess the extent to which ongoing developments in technology already have, and more interestingly, one day might overcome these drawbacks. More specifically, it provides together with this chapter, an elaborate introduction into the issues on which the research described in the later chapters has been founded.

Chapter 3

Meetings and Assisting Technologies

3.1 Introduction

Where science produces information and knowledge about certain phenomena in the world, engineering is the process that leads to the design and the realization of tools and systems that exploit the information about these phenomena for practical human means. The collection of all engineering products, or the consequences of science and engineering, is what frames the concept of technology. Technology has had a significant impact on our daily lives. It has changed the way we travel, the way we spend our free time, the way we learn, and the way we do business. McLuhan (1964) even regarded technology as an extension to the human body.

The revolution in the business world dates back as far as the invention of telegraphy in the 1850s. The invention of the telephone, networked computers, e-mail, and more recent developments in wireless communications and video-conferencing systems have changed businesses dramatically in a sense that they have become much more flexible and efficient. As a consequence team meetings, worker cooperation, and conversations in general have more and more been replaced by email, conference calls, and shared data access. A high speed internet connection, a webcam, a microphone and a few speakers offer employees access to almost all the resources they need.

Technologies can be classified based on various dimensions. One can focus on the technology's ability to act autonomously, on its sensing ability, its reasoning ability, or its acting ability. A light switch, for instance, will typically undergo an external trigger before acting by turning on the light, whereas a radiator might regulate the temperature pro-actively if it able to sense the outside temperature.

The main aim of this chapter is, together with the previous chapter, to function as an introduction to the subject. As subsequent chapters will elaborate on how technology can be used within the meeting domain in order to gain

automatic insights into human meeting behavior in face-to-face settings, it is important to understand the aspects, concepts and issues that play a part. This chapter covers several types of meeting assisting technologies and shows how technology has had, and will have, an impact on meetings and business meetings in particular. Furthermore a variety of possibilities is explored in which technological support can aid meetings, before, during and after they occur. It is shown that technology has altered the notion of a meeting in a way that, instead of physically sharing the same environment, the opportunity to mentally share the same environment has become a more frequent condition for people to interact. Along with this overview it is shown that gaining automatic insight into the human communication process is an important factor to the successful development of future assisting meeting technologies.

Section 3.2 will be concerned with perhaps one of the most outstanding achievements of technology in the meeting domain so far, the ability to interact remotely. Section 3.3 then elaborates on technologies that can provide live meeting assistance. The section focusses specifically on Group Support Systems and Software Agents. The next section, Section 3.4, zooms in on technology that can assist meetings before and after a meeting. In particular browsing technology and meeting scheduling systems are considered.

Instead of having technology brought into the meeting room one can since the first text messaging systems, also have the meetings brought within the technology. These virtual meeting environments, such as Active Worlds (Tatum, 2000) and Second Life (Jones, 2006), have become more and more prevalent. They show an increasing resemblance with face-to-face communication. Section 3.5 goes into more detail about these virtual meetings and discusses the changes in human-human interaction that they bring about.

3.2 Meetings in time and space

As mentioned in Section 2.3.2, the growing array of technologies has developed the very idea of meeting itself. Technology has impacted the way people realize a meeting and how people work together. The aim of all of these technologies is, amongst many other things, to have better meetings, to help teams to work faster and to enhance information sharing and decision making.

The time and space matrix introduced by Johansen (1988) is a matrix made up of four distinct quadrants in which existing meeting technologies can be characterized. This characterization, depicted in Figure 3.1, divides meeting technologies along the axes of time and space. Where in the previous chapter we only considered meetings that took place at one location in a single continuous time interval, this chapter will, in order to give an overview of technological implications, also consider meetings taking place at more than one location and in more than one time interval.

The time dimension has always existed. If people cannot share thoughts at the same moment, one has to communicate one after the other, or asynchronously. This can be unavoidable when participants are located separately

and messages can only be exchanged at very low speeds. On the other hand, if people are at the same location, their attention might be required somewhere else. The solution in both cases is related to the storage and access of information and/or equipment. Technology enabled a shift from a real physical storage place, (e.g. a room, or a bulletin board where participants can leave messages) into a virtual storage place (mail boxes and wiki's).

The foremost benefit of technology so far, however, with respect to meetings, is related to the second dimension: space. The fact that one can have one single meeting, with participants at more than one location can be fully attributed to the developments in technology. Where in the early days messages, or letters, were brought by postal coaches in order to collaborate at a distance, the invention of the telegraph, and later the invention of computerized conferencing[1] has made it possible to transmit messages over long distances in increasingly shorter periods of time. Not only did this decrease the time to inform one another, it facilitated the decision making process and created an extra dimension for the meeting phenomenon. The area of Computer Mediated Communication (CMC) was born and as a result the concept of meeting shifted from physically sharing the same environment into a paradigm where participants could 'meet' as long as they can mentally share the same environment.

What used to be the telegraph, nowadays has evolved into video conferencing systems that are augmented with advanced services such as instant messaging, file transfer and application sharing devices. Even so, the transportation of Morse code and phonetic alphabets evolved into the transportation of more and more communication channels. The possibility for individuals to interact with greater numbers over larger distances and at faster rates than face-to-face communication and the ability to attend meetings remotely has resulted in substantial savings of time and money. Implications of computer mediated communication in relation to human meeting behavior are discussed in Section 3.5.

Given the observations about meetings in space and time, one can identify four meeting classes:

(1) Face-to-face meetings (constrained by space and time). These are structured communication activities in which all participants are physically and simultaneously present.

(2) Physically distributed meetings (constrained by time). These are meetings in which the participants are in different locations but interact in real time.

(3) Temporally distributed meetings (constrained by space). These are meetings in which all participants take part within the confines of the same physical location but are not active at the same time. Examples are senate hearings and corporate interviews.

(4) Temporally and physically distributed meetings (unconstrained by space and time). These are meetings in which the participants do not need to be at the same location and do not need to synchronize the time at which they

[1]The foundations for computerized conferencing, by means of transmitting written messages on computer terminals amongst remote participants, were laid by Murray Turoff and colleagues in the mid 1960s in an effort to upgrade an emergency room of the president of the United States (Hiltz and Turoff, 1978).

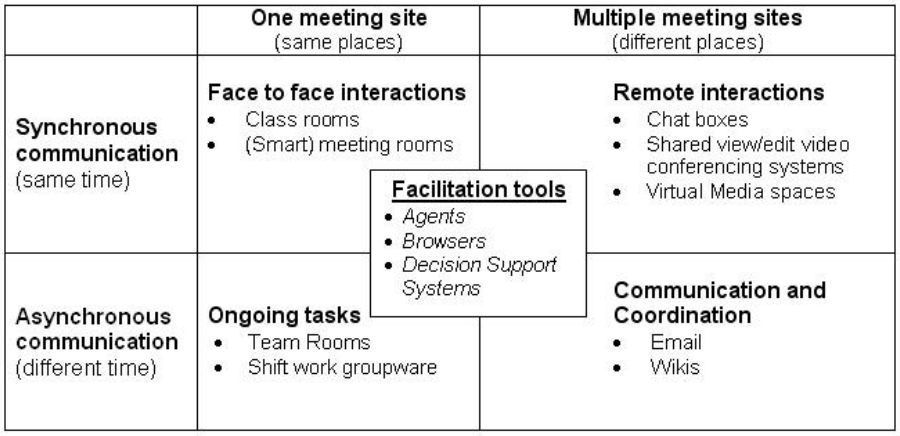


Figure 3.1: An extended version of the Time and Space Matrix

make their meeting contributions. The work done using wikis, voice mail and electronic mail are examples of this class.

3.2.1 Meetings along the Virtuality Continuum

The notion of having a meeting remotely without physically being present in the meeting, but having a representation (e.g. a video stream) there where the actual meeting takes place introduces another dimension that can help to structure the view on meeting support: The virtuality continuum. The concept of the virtuality continuum was introduced by Milgram and Kishino (1994). In this continuum (see Figure 3.2) there is an increasing degree of computer-produced stimuli from left to right.

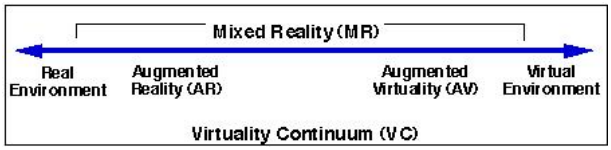


Figure 3.2: The Virtuality Continuum

At the extreme left the environment is completely real, whereas at the extreme right there exist completely (immersive) virtual environments where all stimuli are computer generated. Recasting this in meeting context, we find at the left side face-to-face meetings with real humans in one location and real equipment. The more we move to the right, the more mediated meetings will emerge. At the far right we find the immersive virtual meetings where everything

is virtual: humans are replaced by avatars, the location is a virtual environment and all communication signals are technology mediated. Along with the virtual reality continuum, the notion of sharing the same space evolves from physically sharing the same space to mentally sharing the same space.

The uptake of human computing (See Chapter 1) resulted in more and more face-to-face signals available at the right side of the continuum. Along with the expansion of these CMC related technologies an increasing scala of *facilitation tools* emerged. These tools, that range from passive microphones to complete pro-active systems, have been created to support any meeting irrespective of the time and space constraints and have therefore been placed at the center of Figure 3.1.

A typical example of such a tool is a decision support system that supports the group decision making process. This, and basically any facilitation tool, requires the ability to sense meeting information, reason about this information and possibly act on the sensed information. These requirements are the central theme of the following chapters, but before the methodology is introduced in more detail, an overview of the emergent facilitation tools is given. The first section deals with tools that can provide live meeting support and the second section with tools that can provide support both before and after the meeting.

The floor for these (ambient intelligent) systems potentially can be everywhere along the virtuality continuum. However, when a system needs to change the environment, meaning that the actions that are performed change the way the world is arranged, this will be harder to achieve in the physical world than in a virtual one.

3.3 Real-time Meeting Support

Starting with probably the first meeting ever held by humans, people have looked at techniques and protocols to enhance them. Hotels and resorts even advertise their meeting equipment on their web sites as it (apparently) can be a factor of importance to attract people. On one of these lists I encountered, the following available passive technologies were mentioned (in order of appearance): An internet connection, an instructor podium with a windows-supported computer, a Macintosh computer, or both, plug-in capability for the computer, a CD-ROM, a Zip drive, a Remote/wireless mouse, a ceiling mounted video/data projector, a microphone audio system for multimedia presentation, room lighting controls, a VHS/DVD combo unit, and finally, a telephone. This section will, rather than cover all of these, focus instead on two main types of supportive technologies that can aid the meeting in real-time: group support systems (GSS) and software agents. The focus has been put on these two types because of their dependency on the input side and their large (forecasted) impact on the meeting process on the output side. Both technologies might function as the application domain for the detection algorithms of influence hierarchies and argumentation structures described in the subsequent chapters. The technologies described here function in every quadrant of the time and space matrix and can

(hypothetically) also exist in virtual reality.

3.3.1 Group Support Systems

Group support systems are interactive computer-based environments that support concerted and coordinated team effort toward completion of joint tasks (Nunamaker Jr. et al., 1996). They support alternative, technology enabled, meeting processes that help participants with the formulation of, and search for, solutions to 'problems' listed on the meeting agenda. GSS find their origin in Group Decision Support systems, or GDSS. The 'D' however disappeared as these systems started to support the meeting on the more general levels of information exchange and information presentation (De Vreede et al., 2003). Large displays can, for example, therefore be considered as (part of) a GSS. In general however, GSS systems are designed for finding solutions or creating decisions for problems that have been identified beforehand.

GSS typically combine hardware, software, and network technology to connect participants through terminals to a central server on which several problem resolution tools are available. Examples of such tools are an electronic brainstorming tool, an idea organizer, a topic commenter and a voting support tool. The typical path to problem solving goes through a number of phases, where each tool plays its own role. During the initial 'brainstorming' phase people can anonymously enter on their keyboard all sorts of ideas and possible solutions. In the second phase people (again anonymously) focus on and edit all the ideas generated by the group in the initial phase. Then in the third phase the remaining solutions can be ranked, critically assessed and voted upon. The steps two and three are to be repeated until a final solution is reached. A GSS is usually accompanied by a facilitator who moderates the systems, guides the process and picks the tools that are to be used.

GSS this way provide an opportunity for addressing the aforementioned negative aspects and inefficiencies associated with face-to-face meetings that were shown in Section 2.2.2. Especially the realization of anonymous participation has been put forward as a great advantage (De Vreede et al., 2003; Nunamaker Jr. et al., 1996). Anonymous participation could lead to a reduction of group domination by one or two influential participants and may lead to decreased inhibitions and the reduction of fears for retribution (see also Section 3.5). Furthermore, a GSS allows participants to contribute at the same time. This, in potential, reduces the meeting time as people do not have to withhold their contribution until the others are finished, thereby resulting in the positive side effect that participants can think for themselves, rather than absorbing and remembering the ideas of others. A third advantage is that all generated content is available digitally during and at the end of the meeting. This way information can be better processed and easily stored for later access (see also Section 3.4.2). De Vreede et al. (2003) conclude, after an extensive research on GSS's, that in general they provide added value to the meeting, resulting in higher perceived meeting effectiveness and higher participation levels of the participants. But more important is perhaps the fact that, as (Leven-

thal, 1995) mentions, these systems can be applied in a wide variety of meeting settings as nowadays both mobile and web based versions exist.

Despite the savings and proven increase in efficiency, there obviously exist drawbacks. They are costly, hard to operate and the adoption has proven sometimes problematic, as these systems radically change the way people are used to meeting (Galaczy, 1999). For group negotiations, expert consultations and in situations with a lot of tension, GSS systems have proven counterproductive (cf. Vreede de and Muller (1997)) and it is therefore not strange that Nunamaker Jr. et al. (1995) reports instances in which the use of these systems has been discontinued due to stakeholders' objections.

Mentioned reasons are that anonymity made participants less cooperative and that reaching a decision was therefore harder. For negotiations anonymity also does not work as for investigations about the negotiation space it is important to know who said what. Apart from the anonymity aspect, the fact that ideas become available electronically right away withholds participants from presenting their stances, as they think it will be harder to alter or change them over the rest of the meeting. Participants also indicated that they felt the need to verbally clarify their contributions, especially because a large part of the generated ideas is never worked out due to over expression of ideas. The role of the facilitator has also been mentioned as problematic. It takes a long time before these people are trained and chances are small that they will continue in this job for the rest of their career. Also their role in the process is sometimes mentioned as too influential (see (Briggs et al., 2003)).

3.3.2 Software Agents

An alternative technology that can influence the meeting process and outcome and that will potentially encounter less resistance is the upcoming area of software agents. Software agents are the embodiment of what Hewitt (1977) called *ëactori*, thereby referring to the concept of a self-contained, interactive and concurrently-executing object. This, so called concurrent actor model evolved over the years into 'intelligently' acting software systems that operate alone or in groups in order to fulfill their design objectives. Software agents differ from conventional software in that they can be (semi) autonomous, proactive, and adaptive.

Nowadays, one encounters agents in various disciplines, including e-commerce, business process management, entertainment, manufacturing and, of course, meetings[2]. These meeting agents, or meeting assistants, typically can integrate themselves into their surrounding environment, offer a wide variety of support, and in essence realize the concept of human computing.

Meeting agents have been the topic of research in various projects, with perhaps as most prominent one the Neem Project (Ellis and Barthelmess, 2003; Ellis et al., 2003; Barthelmess and Ellis, 2005). The Neem Project sketches three anthropomorphic meeting assistants that show a wide variety of possible

[2]See Wooldridge and Jennings (1995) for more examples in other domains

applications. These assistants have consistent personalities and well-determined roles. *Kwaku* is a 'virtual' participant that takes care of the organizational aspects of a meeting. He, for instance, reacts to discussions that extend over the pre-allocated period of time by reminding participants that they might want to move on to the next agenda item. *Kwabena* on the other hand is a social facilitator that looks after the participants well-being. He monitors the actions a group would want to undertake at each point in time, such as take a break, switch topics, change the level of detail, or pace of the interaction. He senses the participants' wishes via 'Moodbar' tools on which participants can indicate their desired action. Kwabena subsequently takes the initiative to suggest the course of action (e.g. taking a break). Finally, *Kwesi* is responsible for providing the group with relevant information. This can happen upon the request of one or more participants, but also autonomously, as Kwesi perceives when a certain topic is under discussion for which additional documents are available.

As we will see in the rest of this section and thesis, the Neem dream is about to become reality. Assistants have, for instance, already been developed to greet participants and make them feel at ease (Chen and Perich, 2004), to close the curtains and start projectors once a meeting starts (Oh et al., 2001), to alert participants when someone is calling them (Danninger et al., 2005), to provide feedback and ask relevant questions to stimulate further conversation (Jebara et al., 2000) and even to fulfill the role of a party host who tries to find a safe common topic of conversation for participants (Nakanishi et al., 2004). Niekrasz and Purver (2005) already even described the usage of a shared discourse ontology that could serve as common ground for these sorts of assistants.

The modes of operation for all these systems depend on the abilities to collect information (the sensing ability), on the intelligence to think something about it (the reasoning ability) and the means through which they can influence the meeting (the acting ability). (Notice the relation with Figure 1.1). One could imagine that systems like these need to assess what is going on in a meeting, what the current topic of discussion is, which arguments are used, who the influential people are, who contributes most (least), and perhaps even what the social atmosphere is. Comprehension of the social behavior of the participants and the way groups function are undoubtedly the key factor to success for these agents. The next chapters will further zoom in on this topic and show how a system can be thought to automatically assess aspects of human behavior which, when correctly integrated, provides them with the full potential to aid the meeting processes of the future. The next section will explore opportunities for assistance before and after the meeting.

3.4 Pre and Post Meeting Support

Meetings are more than isolated events. Generally, once a meeting is over, people start to work on the results. Action items are executed, a summary in the form of minutes may be distributed and people start to plan for the next meeting. This so-called meeting cycle (Post et al., 2004) shows on the one hand,

that technology can also assist meetings during the period in between meetings, and on the other hand, that the behavior displayed by humans during a meeting can be influenced by previous meetings. This section goes into detail about two types of technology that can be used in the period between meetings: meeting scheduling systems and meeting browsers. Meeting scheduling systems can facilitate the planning of a meeting and meeting browsers can facilitate the presentation of meeting content and aid *search and retrieval* tasks of meeting content (see Girgensohn et al. (2001); Wellner et al. (2004); Tucker and Whittaker (2005)). Both meeting technologies can also be used during a meeting process (see Post et al. (2007)), although their functioning becomes more apparent in the period between meetings.

3.4.1 Scheduling Systems

The process of meeting preparation can be a very tedious task. Opportunities for technology therefore already emerge in this phase of the process. During the preparation phase insights are obtained about the topics that are to be discussed and the group of people that will participate. Once sufficient clarity is established, the time and location where the meeting will take place are to be settled. Indeed, for all the aspects mentioned above, including the choice of actual participants, technological solution can theoretically be applied. Systems could suggest participants given a pool of employees based on the personalities of the expected other attendants (see the SYMLOG agent as described in Wainer and Braga (2001)) and propose a group size that suits the topics of the agenda (c.f. Padilha and Carletta (2003a)).

State-of-the-art shows that efforts have mostly been limited to the (automatic) scheduling aspect of meeting preparation. With perhaps as most striking success, the nowadays pervasive digital, web-based, calendar systems where people access, organize and optimize their daily activities. Through interconnecting these individual calendars the opportunity emerged to *search* automatically for a time slot that is unoccupied for all of the intended participants. These sorts of electronic meeting planners have been around at least since Wang's alliance calendar was launched in 1984 (cf. Ehrlich (1987)). Ever since, these systems have evolved into more and more useful systems, offering an increasing variety of associated tools. One drawback of these systems is, as already mentioned by Ehrlich (1987), that before one can take full advantage of such systems, the commitment of all members is required. If people refuse to open their private calendar for the system, its forecasted benefits will never be achieved.

There has been some research on agents that schedule meetings, for example in Garrido and Sycara (1995). One of their conclusions is that when agents are given the authority to agree on a meeting time, hiding one's own calendar did not influence the quality of the decisions in terms of the preferences of both agents. In a system described by Berry et al. (2005) agents that function as a *personal assistant* have the ability to negotiate with other personal assistants for a suitable time and location given people's constraints or preferences. Constraint satisfaction approaches have been used by Hassine et al. (2004) to optimally

schedule meetings according to the preferences of all the participants. Furthermore, once the date and location are settled, systems could inform participants about possible changes in the schedule and start to gather the documents to be discussed. Others could prepare the data projector, the light settings and temperature settings of the room and schedule the presentations such as mentioned in Chen et al. (2004).

All of these preference-driven negotiations lead to more and more flexible meeting scheduling. An agent that has enough perception and reasoning capabilities can one day schedule a meeting at a time of which it knows that all participants involved will be in the same building. An implementation described by Oh and Smith (2005) attempts to learn these preferences by 'looking along' with the users for a specific time before starting to contribute suggestions. This learning aspect, that tries to replicate human behavior, is the theme that is central to the next chapters.

3.4.2 Browsers

The preservation of meeting information, also referred to as *group memory*, is due to the volatile nature of meetings gaining increasing attention. Also, people might be interested in things not captured in the notes and hence, as it might take a while to find answers by digging through hard-copy notes, a need exists for technological support.

Tucker and Whittaker (2004) and Tucker and Whittaker (2005) provide overviews of systems grouped into four categories able to browse through (representations of) meetings. The first three categories can be grouped around three classes immediately presenting themselves: Browsers focussing on *audio* (including both presentation and navigation), browsers focussing on *video* and the third class of browsers focussing on meeting *artefacts* such as slides and documents. A fourth, and probably the most useful class of possible browsers, can be grouped around *derived dataforms* providing insights into higher level information such as argumentation structures and influence hierarchies. The AMI JFerret Browser (Wellner et al., 2004) is, for instance, such a browser where several plug-ins potentially can work together in order to distill this higher level information as answers to a specific query.

A very important question is the one Buckingham Shum (1997) mentions: What information is to be captured and preserved, or, what do people want to remember from meetings? In some projects the user requirements elicitation process with respect to meeting browsers has become a research topic in itself (See e.g. (Whittaker, 2005)). The ultimate piece of technology in this sense would be able to answer all questions in a clear and comprehensible manner. A related research area is therefore the automatic generation of meeting summaries (see e.g. (Erol et al., 2003)) as the best summary is one, that encapsulates answers to the most frequently asked questions. The key issue however is, as stated by Palotta et al. (2004), to provide all that are interested with intelligent access to (representations of) meeting information.

As it might be hard for people to express their needs to a system that is

able to tell them something about previous meetings the interface comes into play. Jaimes et al. (2004) describe an implementation of a system that helps users to easily express cues people recall about a particular meeting. On the other hand, Moran et al. (1997) show that, also for the browser domain, people will adapt their way of working based on what they have available in order to increase efficiency. In general three categories of people can be distinguished that might show interest in (parts of) the content or outcome of a meeting: (1) the actual participants, (2) people who did not attend the meeting interested in aspects such as the contributions of a person, or the arguments in favor of or against a specific decision, and (3) analysts who just wish to gather information about meeting processes in general. When focussing on the actual participants, a survey conducted by Banerjee et al. (2005) shows that people, once a meeting is finished, are interested in two kinds of information: (1) descriptions of the interactions among participants and (2) things that involve elements from the meeting domain itself. Similar research has been conducted by Jaimes et al. (2004) and Lisowska (2003).

Once a meeting is over, pro-active agents in the form of assistants could provide selective information about the meeting. Assistants could remind people of commitments and action items they are responsible for. Other assistants might analyze the interaction and produce documents and artifacts that reflect the content of the discussions. The availability of information about what is going on makes it possible to enhance self-awareness and explore ways of providing support to dysfunctional teams from facilitation to training sessions, addressing both the individuals and the group as a whole (cf. Pianesi et al. (2006)). For the provision of information about the interaction amongst participants several techniques have to be developed, able to frame the understanding of what is going on in a meeting setting. This aligns with our observation in the previous section that the understanding of human behavior is more and more becoming a decisive aspect. An example of such a system is CALO's Charter (Kaiser et al., 2004); this suite of agents analyzes interaction during a project planning phase and automatically produces renditions of Gantt Charts sketched by participants on interactive boards. Before going into more detail about how to automatically gain insight into aspects of human behavior, one other dimension of upcoming technology is considered: the developments in virtual reality.

3.5 Meetings and Computer Mediated Communication

Communication at one extreme end of the virtuality continuum, as introduced in Section 3.2.1, is communication in a complete virtual world, where all transferable communication signals are digitally exchanged between the participants. These worlds can vary from a text based chat environment such as IRC [3], up to a complete 3D virtual meeting environment such as the mentioned Active

[3]See: http://www.irc.org/

Worlds and Second Life. Any virtual room, be it a chat based interface, or a virtual 3D environment can be used as a meeting space. Before shifting focus in the next chapters to automatic recognition of higher level meeting phenomena, I elaborate on the impact of technology on the human-human communication process when communicating by means of technological mediation. The focus will be on three main meeting aspects: the group outcome, the group process and the group environment

3.5.1 CMC and Group Outcome

It was Chapanis et al. (1972) who said that the way communication proceeds highly depends on the available communication channels and, indeed, the structure of face to face conversations, for instance, nowadays highly differs from conversations that take place via instant text messaging (cf. Smith et al. (2000)). As the choice of channels affect the group interaction, this in turn, according to McGrath (1984), is also expected to affect the group outcome. Straus and McGrath (1994) stated in this respect, that the more coordination, persuasion, and timing is required from the group with respect to the fulfillment of task (e.g. when the task is judgemental and has to do with values), the more the choice of medium is likely to affect the outcome quality than when correct answers or solutions exist. For an attempt to identify the appropriate set of channels, or modalities, to choose given a particular task and situation we refer to Wainfan and Davis (2004).

So despite the benefits, there seems to be a trade-off and one should not mechanically opt for CMC because of its savings. However, the expectation that the medium will impact the process outcome is not entirely evident; especially not since users of technology have found ways to circumvent the signal omissions. The idea to use smiley's and other emoticons as representations for emotions is an example of how communication is adapted to the medium instead of the other way around. Kiesler et al. (1984) even reports that in 1982, Hiltz and Thuroff already found participants sending computerized screams, hugs and kisses on probably the first text messaging system ever built.

3.5.2 CMC and Group Process

Differences with respect to the form in which communication is realized have been found when comparing conversations on various sorts of media. During video mediated meetings, for example, more formal turn-taking mechanisms are used than during face-to-face meetings (Whittaker and O'Conaill, 1993; Sellen, 1995) and during video mediated meetings, listeners are more hesitant in spontaneously grabbing the floor. Siegel et al. (1986), as well as Straus and McGrath (1994), showed that technological mediation, in comparison to face-to-face communication, might result in seemingly longer lasting discussions, an increase in team member dissatisfaction and even may lead to anormative behavior. Reasons for this anormative distinctive behavior have mostly been attributed to the absence of social and contextual cues that regulate the interaction (especially

when intense discussions and values come into play). Sellen (1995), on the contrary, reports no discernible differences between face-to-face and video mediated conversations, nor between video mediated conversations and audio only, with respect to the number of turns taken per session, the average turn length and the distribution of the turns amongst the participants.

Without the physical presence of others, one is, even when using high quality video channels according to Whittaker and O'Conaill (1993), not able to perceive the other as accurately as in face-to-face communications. The reduction of eye contact and gaze signals, for example, are renowned for their impact on communication as this reduces the ability to create a joint reference to events and external objects (c.f.(Kendon, 1967; Argyle et al., 1973; Vertegaal, 1998; Heylen, 2006)). The lack of information about the other participants according to Kiesler et al. (1984) and Siegel et al. (1986) could lead to feelings of reduced *presence* and a sense of anonymity, which are grounds for the decrease of inhibitions and the reduction of fears of retribution and rejection which, in turn, could explain the anormative behavior. Anonymity, on the other hand can result in more equal participation and an increased task orientation, especially if interpersonal information such as status aspects (e.g. age and background) remain unavailable (Short et al., 1976).

3.5.3 CMC and Group Environment

Presence is determined by the richness of the media, and the abilities it affords to the user. An increased sense of presence leads to enhanced perception of others and increasing possibilities to express oneself (see Short et al. (1976); Lombard and Ditton (1997)). Fisher et al. (1986) in this respect state that the possible ways to communicate increases along with the number of modalities in which one can express oneself. Whittaker (2002) expected that the more face-to-face communication channels were supported by technology, the more of the above mentioned differences in human meeting behavior would disappear. This so-called bandwidth hypothesis, has however until now not been proven. Chapanis et al. (1972) and Short et al. (1976), for example, found that adding a visual mode to speech does not necessarily increase the communication efficiency and that channel combinations including speech were always more efficient than those without.

Whittaker (2002) mentions that as soon as social cueing aspects become critical, the provision of visual personal information such as video signals yield better results for tasks that require access to personal information, such as getting to know each other, than just voice or text signals. The question how to determine the appropriate set of cues for a specific task seems to be the main challenge to be resolved.

The communication towards the right end of the virtuality continuum is inherently confronted with issues resulting from a leaner medium in a sense that not all social cues one encounters in face-to-face interaction are conveyed. Greenhalgh and Benford (1995) were among the first developers of a virtual reality teleconferencing system where participants could participate in a virtual

meeting whilst wearing a head mounted display. The system called MASSIVE provided globally distributed people the opportunity to have a meeting in a 3D environment that was represented at all sides. In comparison to face-to-face meetings they mention for their virtual meetings: limited peripheral awareness (due to the HMD), lack of engagement, depersonalization and a decreased feeling of presence. Furthermore, and notably interesting, a flattened meeting hierarchy is mentioned.

This raises an interesting hypothesis, in a sense that it appears that if not all communication signals, as expressed in face-to-face meetings, are being transferred, participants seem to lose control of the situation. The image participants have about the other participants is becoming increasingly incomplete as more and more signals are omitted. This aligns with remarks that aspects such as affiliation, grounding, intimacy and more recently rapport, have been described as an important pre-requisite for successful communication (Argyle and Dean, 1965; Traum, 1994).

This potentially results in two things. In the first place, one needs to put in extra effort to find an appropriate form to encode a message for transfer, which in turn runs a higher than normal risk of being wrongly decoded by the recipients. For the sender to check for correct understanding he again is confronted with the sub-optimal set of signals, etcetera. The second, and perhaps consequential results of the first is that, whenever the communication medium becomes leaner, the exertion of power becomes more and more difficult due to the lack of control. One does not experience the same process as the other participants, especially not when people are sharing nothing but a text interface. There can be different conditions and distractive events going on at the remote sites that are beyond the control of the others.

Both aspects could explain the reduced feeling of presence, as well as the growing sense of anonymity as a consequence of the lack of control. The lack of control, in turn reduces the chances for retribution and the aforementioned anormative behavior emerges.

To overcome these drawbacks humans compensate for the technological deficiencies by adapting to the channels available and exploiting the technological benefits. One can change the perception of others according to ones own preferences by means of altering the local representation of the virtual environment in which the signals from the others are perceived. Backgrounds from chat boxes, for instance, can be altered and video streams used for teleconferences can be optimally chosen. Augmented reality techniques can even stress, highlight and hide signals according to ones preferences (see e.g. (Barakonyi et al., 2003)). The other way around, one can steer the perception of oneself remotely by, for example, choosing a preferred representation in a virtual world. A person's face can in this way be represented with the face of a pop star, or a wild looking dragon on an internet forum. Bailenson et al. (2004) even mention teleconferencing systems where the behavior of the participants is modified before being sent out. Both ways taken together do not just show how humans use the technology to have the meeting according to ones preferences, they underline the importance of control over the situation.

A trend in ongoing technological developments is, on the other hand, that virtual worlds are increasingly equipped with signals encountered in face-to-face interaction. Apart from head mounted displays that provide enclosed views of a shared 3D virtual world, spatial audio mediation (see e.g. Rodenstein and Donath (2000) for a 2D and Aoki et al. (2003) for a 3D version) and haptic interfaces (Mark et al., 1996), that enable the touching of objects, are becoming more prevalent. As a consequence, from one side the behavior of the other might be even harder to control as signals can be increasingly manipulated, on the other side, the more signals available, the more interaction starts to resemble face-to-face communication. For more information about trends and developments see Section 8.2.

If and how all the above mentioned aspects impact the meeting remains to be investigated. A closer investigation of these issues will be the subject of on-going research in the AMIDA project[4] and falls beyond the scope of this thesis.

The virtual meeting environments, however, offer a perfect arena to introduce autonomous agents that have the same communicative channels at their disposal as the human participants. One can equip a virtual meeting world with software agents that themselves can be simulating a meeting participant. The chatbot Eliza (Weisenbaum, 1966) was probably the first trial to realize this in a text based environment. Software systems nowadays have lifelike embodiments, they are able to show intelligence, express emotions and are equipped with skills to interact with human users (see e.g. Gratch et al. (2002) and Traum and Rickel (2002)). Existing work has already shown that people can be influenced in their behavior as well as their assessment of a situation through the presence and the behavior of these agents, even if the participants know that the agents are not representing a real human (Pertaub et al., 2002; DiMicco, 2004). In the near future these systems could become totally indistinguishable from real human representations: one day they even might have the potential to fully replace a meeting chairman (See Section 8.3). Developments in this spectrum grow along with state of advanced recognition technologies for human-human interaction, the central topic of the remainder of this thesis.

[4] www.amidaproject.org

Chapter 4

Corpus Based Interaction Research

4.1 Introduction

Behavior refers to the actions or reactions of an object or organism, usually in relation to the environment. Behavior can be intelligent, social, and inappropriate. Behavioral actions and the processes involved in their formation and modification have long been known to be conditioned by the types of situations and the experience encountered by animals and individuals along the course of their lives (see Thomans (1927)). Social behavior is behavior that is directed at people, it is an advanced sort of behavior that one typically finds in meetings.

When humans behave socially, or more specifically, when humans interact, they use their natural skills to sense and interpret signals in the environment in a way that specific behavioral responses result. The ability to recognize these behavioral responses and to learn their association to the context in which they occur are critical for an organism's survival. Sinha (2002), for instance, described this skill as a prerequisite for foraging, danger avoidance and mate selection.

Associations and patterns that relate cause and event have played an important role throughout the history of humankind. Where initially hunters found patterns in animal migration behavior, people over the course of history developed tactics and tools to benefit from these findings (See also Section 2.3.1). The research area that deals with the automatic detection of these patterns is the area of pattern recognition. Pattern recognition, amongst others, studies how machines can observe the environment, how machines can learn to distinguish patterns of interest, and how machines can make sound as well as reasonable decisions and inferences.

Recognition and remembrance of behavioral regularities and patterns identify opportunities, and can be turned into new insights, a competitive advantage, and a profitable business. In any social encounter, including the meetings of ev-

eryday life, every living person, according to Goffman (1955), displays both consciously and unconsciously a pattern of verbal and nonverbal behavior and thereby, when recognized, reveals his view of the situation and shows information about the internal evaluation of the other participants. The main challenge here of course is, to obtain the ability to recognize this pattern, to identify its regularities and to find the opportunities for exploitation. Even more, when this can be achieved automatically, a whole new era for human computing applications emerges (See Section 3.3.2).

This chapter describes the methodology known as corpus based research. The methodology is applied as an attempt to semi-automatically distill the higher level meeting phenomena of **influence hierarchy** and **argumentation structure**. Central to the methodology stands a *corpus*. A corpus is a collection of recorded signals that represent a preferably representative sample of a particular phenomenon, such as in our case four person meetings. A corpus embodies a research domain and it generally enables the validation of domain related rules and hypotheses on empirical grounds, as well as that it provides the opportunity for scientific explorations and hypothesis formulation.

As a corpus typically contains mark-up or annotations that signify occurrences of particular phenomena, it can this way be used to check for the coexistence of certain phenomena within particular contexts and for the correlation of particular signals and events in a (semi-)automatic manner. In a corpus that contains just data such as text, one could for example extract word combinations to either create a model that predicts the next word on any word given word from the text, or to validate such a model in terms of correct predictions. However, if this same corpus also contains a Part-of-Speech tag (such as 'Noun' or 'Verb') for each word, models can be built that predict the Part-of-Speech tag given a word (See e.g. Brants et al. (2003)). These models that explicate patterns in the data, and that transform data into information, can in turn also be validated.

So, as long as certain information is explicitly included in a corpus, the methodology enables algorithms to learn how the information can be retrieved from other information and data that are encapsulated and it, at the same time, allows for the validation of the models that describe the information.

As human interaction research is not a new field of study by itself, over the years many problems and solutions to the problems have already been identified. Section 4.2 provides a little background on historic approaches and outlines the main hurdles on the road. Section 4.3 then provides the framework of the methodology. It identifies the main related components and it shows which steps need to be taken before the framework can be applied. It will become apparent that a predefined sufficiently detailed model needs to be available that describes the phenomena one wants to recognize, or replicate. How to create such a model, or annotation schema, and what aspects have to be considered are subject of Section 4.4. The chapter finishes with a description of the techniques and algorithms required for replication and automatic application of the model on unseen data in Section 4.5.

4.2 After Models of Interactions

Although the automatic observation and interpretation of human interactions has only recently become an application domain for human computing research, it is not a new field of study. Social psychologists, for example, have been actively engaged in the development of explanatory (Smith, 1942) and descriptive (Bales, 1950) models of behavioral patterns for over 60 years. All of the developed theories of group behavior and interaction research can, when operationalized, potentially be used for the creation of socially aware systems. They provide valuable insights that can be exploited for the creation of the quantitative and mathematical models suited for machine perception.

One of those early social psychological methodologies that aimed to provide structured insights into the structure and functioning of a small group was the Interaction Process Analysis, or IPA, devised by Bales (1950). IPA, later developed into the method of Systematic Multiple Level Observations of Groups, or SYMLOG (Bales and Cohen, 1979) categorizes the "unit of speech or process", or the behavioral actions of humans in social settings in a *fixed number of predefined categories*, or coding schema. Examples of these categories are 'Shows Solidarity', 'Asks for Opinion', 'Shows Tension', 'Agrees' and 'Shows Antagonism'. Over the course of an encounter, each of these units of speech were put into one of these categories, once the meeting was over the resulting distributions of group member actions were used to describe differences in the roles of the group members. This system, this way provided, amongst others, the possibility to compare members with each other, and to group the members into those that focus on the task and those that focus on the process. Similarly, the hunter from the previous paragraph could have described the migrating behavior of the animals in terms of whether the animals would stay or leave their feeding grounds.

The differentiation of phenomena makes it possible to categorize, and perhaps even order them. Categorization predicts a relationship between 'subjects', or 'samples', and knowledge domains. These relationships in turn can aid inferences and predictions central to the subject or sample. The differentiation between edible and inedible animals, or edible and inedible fruits, could for the hunter be a matter of life and death and similarly, within the meeting domain if one, for example, can choose between a task oriented leader and an emotionally oriented leader if a meeting has to stay rational a differentiation undoubtedly pays off.

This linkage of causes and events underlines the need, not just to differentiate the phenomena into distinctive classes, but also to know their constituents as well as their possible implications and consequences. The set of constituents, or *features*, that was used by Bales's IPA served the differentiation of the participants into functional categories. Similarly, the hunter, for instance, could have considered the smell or the color of the available food in order to predict its edibility. Features are, to put it in a more general sense, the individual measurable properties of the phenomena observed. A particular distribution of *value(s)* pertaining to the set of available features result phenomena to correspond to

particular classes. To give an example: if an object is cold and transparent, and one can choose between an ice cube and a sugar cube, it is more likely to be the ice cube. The extent to which the features discriminate the classes from one another are therefore generally considered the key to successful classification.

The IPA categories are faced however, with the trade-off to, on one hand provide sufficiently numerous categories, or features, to cover for the distinctions in the perception of the unit of speech by the observer, whilst on the other hand, the number of categories was to be small enough to allow for a ready and comprehensible overview for the observer. This seems to pose an unsolvable problem. An even more fundamental issue in relation to IPA was brought up by, Harvey Sachs, an influential American sociologist, who in one of his first lectures on Conversations noted that there is no reason to suppose that "the observer has it right"[1] (See Sacks (1992)). According to his beliefs, one should take the smallest pieces of human behavior and try to collect those that look alike before saying something about it in a greater context. This, so called example based approach to conversational analysis (CA) starts, in contrast to the sort of top-down IPA approach, by very precise examination of single cases and progresses in a bottom-up manner to the discovery of the structures that organize them.

Although the example based approach did not tell anything about how often the investigated phenomena actually occur in real life, and Bales at least provided a model that described what he was after, Sacks was right in a sense that all observations are subject to errors. Sacks's approach as a result has been of great influence over the years, as the common tactic to describe behavior has long been to examine one isolated behavioral signal at a time[2]. It increased the insight into and knowledge about social behavior, but it has been of limited usefulness in efforts to describe the behavior one encounters in everyday life. In current everyday practice, if one is willing to model aspects of interaction research, it does not really matter which approach is taken, as long as a model is constructed that allows for systematic observations. The place where one starts is often dependent on the researchers preferences, the goal of the enterprize and/or the environmental conditions that come along. As will be described in Section 4.5.1, both of the Social Psychologically inspired approaches to model human interactions now have their computational counterparts: The Balesian tradition in the area of concept learning and the Sacksian tradition in the area of data mining.

4.2.1 Statistical inference

Section 2.3.3 used descriptive statistical methods to summarize and describe certain aspects that belong to the concept of a meeting and that create grounds for inter meeting comparisons. The average meeting duration and the mean

[1] See Perakyla (2004) for an elaborate comparison between Bales's IPA and Sacks's Conversational Analysis approach

[2] See, e.g. Argyle and Dean (1965) on distance, Kendon (1967) on gaze, and Sachs et al. (1974) on turn-taking.

number of meeting participants, for instance, provide, together with the associated standard deviations and distributions, an image that allows for this. To further examine dependencies between the charted aspects inferential statistical methods can be used.

Inferential statistics model patterns in data in a way that one, for example, can test the hypothesis if a meeting generally lasts longer when more participants are present, or can predict the meeting duration when the number of participants is known. Statistical analysis can thus reveal if two aspects are correlated, that is, if they tend to vary together as if they are connected. An important consequence of this is that these techniques can in this way aid us with the identification of more easily automatically detectable meeting aspects that correlate to higher level meeting phenomena.

However, when wishing to make descriptions and inferences one is usually in a situation that just a part of all the data available can be studied. This limitation poses another problem on our quest, namely the extent to which the chosen subset is representative for the set as a whole. Especially in observational and experimental settings representativeness can become an issue[3]. If the examined subset does not generalize, neither will the distilled models of descriptions and inferences. One could speak in this case of *overfitting* the model(s) on the (training) data.

All in all this section identified three challenges that are to be resolved on the way to higher level meeting phenomenon detection: (1) A coding scheme needs to be devised that systematically represents the phenomenon in sufficient but not too great detail in order facilitate the observations. (2) The coding scheme should be mapped correctly onto the data before inferential statistics can be made. (3) The data should be representative, as the distilled inferences and descriptions should generalize. The next section will describe the methodology that throughout the rest of this thesis has been used to deal with these challenges.

4.3 Corpus Based Research for Human Human Interaction

Sixty years have passed since Bales's IPA methodology was published. Ever since that moment, the increasing availability of computers immediately led to the creation of (initially text)corpora in electronic form that could be searched automatically for a variety of features (See Ide (2004)). Algorithms were applied to distill frequencies, distributional characteristics, and other descriptive as well as inferential statistical measures.

An example of one of the first large corpora that appeared was the Brown Corpus in 1967 (Kucera and Francis, 1967). This corpus contained over 1 million words tagged, or annotated, with part-of-speech information. Along with the rapid increase of computational power in the 1980's and the increase in data

[3]See Vissers et al. (2001) for an interesting discussion

storage facilities corpora appeared that were increasingly larger and contained an increasing variety of signals, or modalities. Corpora that consist of more than one modality are called Multi-modal corpora. Multi-modal corpora typically contain recordings from various sensors, such as video and audio, that are all annotated with large varieties of phenomena. As outlined in Chapter 1, the era of human computing provides more and more opportunities for the support of human activities. Corpus based research facilitates the creation of models that are able to interpret the sensed environment. Where the early corpora were used to train models for syntactic parsing, the AMI corpus, that was introduced in Section 1.3.1, has specifically been created for the investigation of patterns of human-human interaction in four person meetings aiming for development of theories of human communication as well as the development of supportive meeting technologies.

Central to the task here is the automatic assessment of higher level meeting phenomena that can be obtained, preferably at low cost and in real time. A corpus contains stored, non-volatile, data that can be empirically investigated and augmented with all sorts of annotations. The plethora of available signals seems an appropriate starting point for investigations into the feasibility of the task. Figure 4.1 gives a schematic representation of the aspects involved into corpus based research.

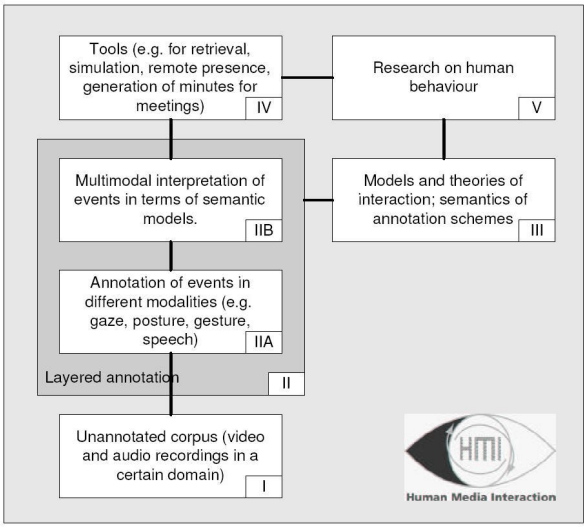


Figure 4.1: The corpus based research framework

The process of corpus based research starts with the corpus of interaction recordings (Box I), on which manual or *automatic* recognition processes apply a predefined coding scheme. This results in recorded observations, or annotations, that systematically describe the data (Box II). These annotations can be used by

a variety of tools (Box IV) that can provide real time support and/or provide hindsight insights into the meeting process (See Chapter 3). The obtained insights can be used by social scientists to further investigate the corpus data (Box V), for example, by means of statistical analysis. The analysis in turn could lead to the development of new techniques and tools and also (potentially) result in new coding schemas and new theories (Box III).

Given this model, the three issues of concern mentioned in the previous Section can be positioned. The representative data that are subject to investigation resides in Box I. The coding schemas that are to be applied on the data are part of Box III and the mapping from data to observation itself takes place on the line between Box I and Box II.

The term *layered annotations*, as Box II is called, relates to the level of interpretation of the observations. Box II is therefore divided into two subboxes that separate the objective, directly observable annotations from the more semantically oriented annotations, that is, those that require interpretation[4]. To give an example: what in Box IIA would be called a 'hand-raising' event, could in Box IIB be called a 'request' or a 'vote'. The layering becomes apparent as for a 'request' to be observed, or recognized, the detection of a hand-raising event can be of great value. Or to put it in terms of the previous section: 'hand-raising' could serve as a *feature* for the observation of a request and fulfills in this way a similar function as the features that allow the IPA system to detect (and to discern) task and process oriented members. This implies, in the first place, that the observations of the phenomena that we aim to detect will eventually end up in Box IIB, and second that the detection of specific observations, or a specific *layer of annotation* is dependent on other layer(s) of annotation. The annotation layers strongly relate the the layers of interpretation mentioned in the 'Recognition' box of Figure 1.1. So if we want to automatically detect higher level meeting phenomena, we in essence want to end up with interpreted annotations that are inferred from more objective (layers of) annotations.

All in all, the procedure turns out as follows. Given a corpus that contains the phenomenon of interest, an annotation schema has to be devised that describes this phenomenon. This schema has then to be applied on the data after which correlated, more objective, and more easily detectable annotations, or features, are to be identified that in combination can lead to the replication of the initial annotation schema.

Before it is explained in more detail how to create an annotation schema, and how to find a set of features that can replicate such a schema, this section first continues by providing more elaborate information on each of the boxes involved.

Box I: Corpus

Research on multimodal interaction often uses a corpus of audio and video recordings. In general, a corpus should be representative of the domain, be

[4]This distinction is more or less comparable to that of Bakeman and Gottman (1997) who discern between a physically and socially based coding scheme.

large enough to do relevant research and be accessible. The data residing in a corpus can usually be distinguished as a collection of signals (captured data from sensors) and annotations (signal interpretations describing the data). A corpus should furthermore be extensible, in a sense that annotations can be added on top of one another.

For the signal collection many projects in the area of interaction research use smart rooms. These smart rooms are equipped with a range of sensors (visual, aural and other types) that allow, often detailed, capturing of the interactions in the room. Audio is in most cases recorded with lapel microphones, binaural mannikins or microphone arrays. There are video cameras that capture the whole meeting room, or parts of it from different viewpoints, individual participants or even close-ups of their faces. Documents pertaining to the meeting are collected or captured and even writing is recorded using whiteboards, or smart pens. Examples of existing corpora that contain meeting recordings more or less similar to the AMI corpus are those used in the Meeting Room project at Carnegie Mellon University [5], in the Meeting Recorder Project at ICSI[6] and in the M4 project [7].

Box II: Layered Annotation

To study signals that have been captured, annotation schemes have to be designed. Annotations are what some call 'metadata', and metadata generally have the role to apply *descriptions* of properties and content to data. These descriptions can apply to any phenomena that can be observed. They can be applied at an individual level or at a group level, they can be applied to the environment and even to the investigated process itself. Preferably these descriptions are based on theoretical models and chosen because they are useful for a particular domain of application. As stated, annotations as well as the theoretical models can describe meetings at different levels and ultimately all annotations can be created automatically, reliably and in realtime. The annotations on a meeting corpus might include transcriptions of the speech, the names of meeting participants, their speech acts, the gestures that are made, who the speaker is, the head and gaze orientations, the addressees of the speaker, the focus of attention of the participants or the group, the displayed emotions, and current topic. The actual design process of an annotation schema is discussed in Section 4.4.

Box III: Models and Semantics

To enable the interpretations of annotations from box IIB, there exists a need for models of human interaction. There is a large variety of examples: models of the dependencies between group behavior and leadership style (see e.g. Hersey and Blanchard (1988)), a model of the rhetorical relations between utterances (see e.g. Kunz and Rittel (1970b)), an agent model incorporating emotions (see

[5] http://www.is.cs.cmu.edu/meeting_room/
[6] http://www.nist.gov/speech/test_beds/mr_proj/
[7] http://www.m4project.org

e.g. Pelachaud and Bilvi (2003)) and many more. All models generally result from their own research objective or applications. The models can be derived from corpus investigations and, for example, result in hypotheses of people's intentions in relation to certain behaviors. As described, a corpus in turn also allows for the validation of these models. [EXPAND, relate to figure in chapter 2.. all sorts of models]

Box IV: Tools

Corpus Based Research might lead to the development of tools for supporting the phenomena embedded in the corpus. The functionality of the tools depends on their *level of understanding* of the environment. This level of understanding is determined by the extent to which the tool is able to create, or access annotations of a certain level. The direct provision of certain annotations to meeting participants has already proven to be beneficial by itself (DiMicco et al., 2004).

In the context of meetings, these tools are typically useful for end users and basically relate to the tools described in Chapter 3 such as a meeting browser, a minute generator, a remote meeting assistant. These are all tools that are useful in the domain that is represented in the corpus. Other tools that can be developed are tools that aid (parts of) processes in the framework of corpus based research, such as tools that support the manual annotation process and simulation tools that can be used for the validation of hypotheses.

Tools that are based on theoretical models and algorithms that obtain some of the annotations automatically (from other annotations) generally cannot do this with a hundred percent accuracy. This is especially true for the extraction of the more higher level annotations, as these are generally based on error prone recognition techniques (c.f. those used in speech recognition and image processing). It must be said that this goes for humans as well. Humans also find it more difficult to describe facial expressions in terms of emotions in comparison to, for example, the tagging of words with the appropriate part of speech. So, generally the following observation can be made : the more the observations rely on interpretation, the more the descriptions of the observations will diverge.

Box V: Human Behavior

Research on human behavior, for example social psychology, provides an insight into human interaction patterns and their components. These insights help to discover more about human nature and consequently satisfy particular human needs that provide opportunities to further develop themselves.

The emergence of social patterns, such as described in Section 2.3 form the basis for automatic analysis and for the retrieval of components. The corpus can be analyzed by means of tools to discover regularities in annotated human behavior and construct corresponding models and hypotheses. These models in turn can be evaluated, for example, by using the corpus itself, but also by means of simulations and user studies (See Section 4.4.3). The next section will go into more detail on how to distill these models and annotation schemas.

4.4 Modelling Data

As humans behave socially, the detection of social signals could lead to systems that can listen in to interactions and define their action on this (Pentland, 2005). That is what Human Computing is about: systems that respond to events that occur naturally in everyday human life. As system responses generally are a result of internal models that describe how to transform input signals into output, before thinking of an output in the area of Human Computing a system some way has to be able to assess the potentially relevant aspects of its interacting inhabitants. Bakeman and Gottman (1997) described these models as follows: " These models, are the lens in which one has chosen to view the world [..] If that lens is thoughtfully constructed and well formed, a clearer view of the world emerges". This means that systems can increase their perception of the world when equipped with the appropriate models. Increased perception in turn can yield increased output, in a sense that a system is better equipped to fulfill certain user desires, be it either the fulfilment of a specific task, or the way that humans interact with the system itself.

Most of these models are often difficult to derive. This is mainly due to the fact that these models define the internal architecture of a system. The development of such an architecture is a hard problem, or a design task, with generally more than one solution. Systems that should understand aspects of human behavior in the meeting domain are an even more difficult enterprize; they require the quantification and interpretation of social signals that can be exposed in several modalities and communication channels at one time. A model in the end generally results from the careful analysis of a corpus and its annotations, or through training over time in real life. This section focusses on the creation of models by means of a corpus. This was done in the first place because within the AMI project, a suitable corpus for our domain was present, and in the second place, because real life training requires continuous feedback to the system over a certain, and usually quite extensive, period of time before a system, if at all, becomes able to translate the appropriate set of input features into a desirable output.

4.4.1 Schema Creation

For human computing applications, the models, or coding schemas are nowadays often inspired by social psychological hypotheses that try to describe human-human interaction. The model from Bales, as described in Section 4.2 was able to distinguish between more task-based and the more process-based participants whilst given a set of features that were to be recorded by the observers. He aimed to prove this way that face-to-face interaction contains formal similarities that occur irrespective of the individual participants and their locations. The resulting distinction that he was able to deduce fulfilled his goal of creating operational distinctive variables that are general enough to be applied to a large variety of small groups.

A legitimate question in this context, at least since Bales's finding, concerns

the interest of the researcher, or the goal of the system: What to observe, and why? Bakeman and Gottman (1997) state in this respect that if research questions are clearly stated, it is easier to determine the distinctions a coding scheme should make. A model, or coding scheme, should hence be created in order to fulfill a certain need; be it either answers to the question of the researcher, or as in case of a system, to fulfill part of its goals. Any resulting model, to put it more generally, should make sensible and interpretable distinctions from the data.

To initially determine the correctness of a model, and its associated fit on the data, the often elaborate process of model application awaits. As no algorithms have been trained on the data for the automatic applications, this usually needs to be done manually by means of annotation tools. But if, eventually, a particular model can be successfully applied to the data, the resulting annotations contain useful information for a variety of goals and applications.

4.4.2 Annotations

Annotations are used to codify judgments of observers in relation to an annotation model or schema. They are the tangible result that captures, organizes, and conveys observed information in a structured manner. As mentioned, these annotations can be used for a number of tasks. They can be used to evaluate hypotheses in the area of social psychology, as examples for machine learning techniques that strive for automatic model application on unseen data, and for the validation and re-design of the annotation schemas themselves.

If all observations that are made on the data can unambiguously be classified into one of the predefined schema categories, one could say that the model perfectly fits the data. However, before one can really be sure of this, it is important to be aware of the other two interwoven challenges identified in Section 4.2: The data should be representative, and the observers should know how to apply the model.

To be able to accurately apply an annotation scheme, observers should make judgements about what they observe. This is not always a trivial task. Especially not if the observations require interpretation. To observe, for instance, that someone has 'the intention to ask a question', or that he or she expresses 'a certain emotion' largely depends on a subjective interpretation. Observations that require interpretation rely on more than the knowledge of how to apply an annotation schema. Performing adequate judgements requires observers to understand the 'culture' of the observed interaction and to possess a certain social sensitivity that includes the ability to empathize with the observed interacting subjects. All of these requirements are in line with Sacks statement that there are no reasons to assume that 'the observers see it right' and that observers are more likely to disagree about observations that rely on interpretations. Difficulties are, for example, indeed reported in the areas of emotion detection (See e.g. Steidl et al. (2005) and Batliner et al. (2006)), and intention related discourse tagging (See e.g. Nomoto and Matsumoto (1999)).

Agreement about observations between observers makes it easier to infer

conclusions from the data, or to quote Bales: "We consider ourselves fortunate when we have roughly comparable rates of incidence of a series of phenomena .. When these rates are based on data gathered in a comparable way and conform standard definitions, we are able to make more definite comparisons" (Bales et al., 1951). Thus a high agreement between observers means that observers highly agree on the chosen categories from the annotation schema for particular sections of the observed data. A high agreement is beneficial as the observations now generalize across observers and become more easily reproducible (Cohen, 1960). However, there is a trade-off here between the amount of training that is required for the observers and the desired level of agreement. The more training is needed for the observers, the harder it will be for others to apply the same set of categories with any assurance of obtaining similar results (see (Bales et al., 1951)). The quality of the annotations in terms of agreement is quite often assessed (e.g. by means of the κ measure (Cohen, 1960)). This measurement shows the level of agreement between two annotators corrected for agreement by chance. The issue that arises is that the establishment of the presumed truth can be an endless discussion (See e.g. Bakeman and Gottman (1997)). If both annotators were wrong, the agreement can still be very high. We stay away from this discussion.

Relevant here is the question what one really wants from the data. Does one want to deduce algorithms that can apply a generalized realization of the annotation schema, or does a version that replicates one individual annotator suffice. Especially for annotations with a low inter-annotator agreement, the question is if the model is to be blamed, or if humans just will not agree due to their innate cultural differences. In the case where humans beforehand are likely to disagree, all their observations can be defendable, or to put it somewhat differently: there could be more than one correct observation.

The way the annotations are created, as well as some quality aspects, such as annotation consistency, are relevant for the explanation of algorithm behavior that has been trained on these annotations. How we dealt with this is described in Section 5.4.1 for our efforts to replicate dominance rankings and Section 6.3.4 for our efforts to replicate argument structures.

Many large projects face the challenge of manually annotating a large amount of data for various modalities. The process of creating the annotations by itself is, even without focussing on the training of the observers and reliability of the resulting annotations, a tedious and expensive task. Annotating a stretch of video with not-too-complicated aspects may take ten times the duration of that video. Shriberg et al. (2004) report an efficiency of 18xRT (18 times the signal duration is spent on annotating) on annotating dialogue act boundaries, dialogue act types and associated adjacency pairs on meeting recordings. Simple manual transcription of speech usually takes 10xRT. For more complicated speech transcription such as prosody 100-200xRT has been reported (See Syrdal et al. (2001)). The cost of syntactic annotation of text (PoS tagging and annotating syntactic structure and labels for nodes and edges) may run to an average of 50 seconds per sentence with an average sentence length of 17.5 tokens (cf. Brants et al. (2003), which describes syntactic annotation of a German

newspaper corpus).

If annotations have to be performed manually, one can develop tools that allow for the efficient creation of annotations. Within these tools, knowledge about the phenomena that are to be annotated can be embedded. This embedding allows tools, for example, to suggest annotations, to limit choices, or to pre-fill values of attributes. Efficient annotation interfaces as well as trained annotators might help in this respect, but apart from interface improvements in order to increase the annotation efficiency some automatic annotation procedures can provide support for manual annotation. These kinds of semi-automatic annotation techniques are already being applied for audio transcriptions and video segmentation. Human annotators now only have to correct the automatically detected boundaries. This manual correction is much faster than full manual annotation (Syrdal et al., 2001).

For more information about annotations and issues related to their obtainment see Reidsma (2009).

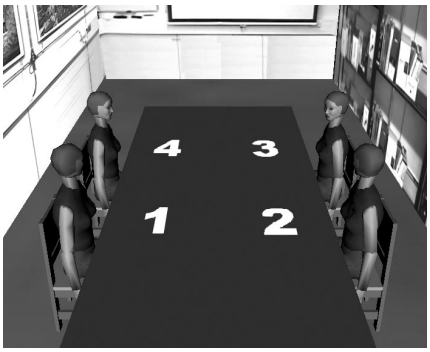
4.4.3 Schema Validation

Annotation schemas can be evaluated in order to be improved. These improvements can sometimes be necessary to realize an easier schema application for the observers, or a better fit with the data. This can happen in case, where particular categories that could describe the observations are missing, or if some are indistinguishable, that is, that they overlap.

An intuitive starting point is a critical consideration of the initial annotations that are produced by the annotators whilst applying the schema that is under discussion. Confusion matrices generated from annotations by various observers and/or algorithms can provide valuable insights in this respect.

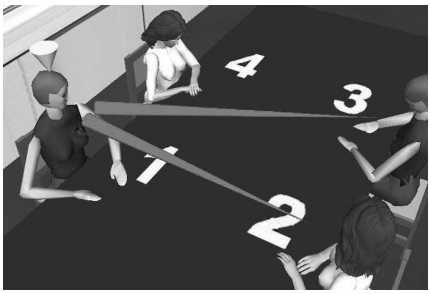
On the other hand, the applied annotations can be used in simulation environments to see how well they fulfill the goals of the designers. See for example the work of Padilha and Carletta (2003b) where certain mechanisms for turn taking in small group discussions are examined by comparing certain models of floor patterns with patterns observed in real life. I will elaborate a bit on virtual simulation environments.

State of the art in computer graphics and embodied conversational agents allows the creation of *Virtual Meeting Rooms* (VMRs), virtual replicas of real meeting rooms (See Figure 4.2(a)). Simulation of a meeting in such an environment, may, for instance, involve the virtual replay of signals and annotations. The information displayed can be both directly obtained from recordings of behaviors in real meetings (e.g. tracking of head or body movements, voice), and also stem from manual and/or machine generated annotations. Where replay of signals allows for closer examination, the replay of signals in a virtual world, also allows for the replay from specific viewpoints, such as the viewing perspective of one of the participants. This might give researchers a unique perspective on the behavior of meeting participants and provide new insights that can lead to improvements as well as the construction of certain hypothesized models, including annotation schemas.

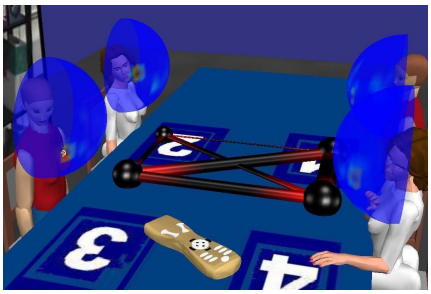


(a) VMR central view

(b) VMR view seen through the eyes of a participant



(c) VMR extended view with visualization of head orientation, body pose, speaker and addressees



(d) VMR extended with derived gaze and influence information

Figure 4.2: Signal replay in a Virtual Meeting Room

A virtual environment, furthermore, provides the opportunity of signal and annotation control. This means that any combination of recorded signals and annotations that a corpus contains can be closely examined for evaluation purposes (see Loomis et al. (1999)). One could think of the evaluation of descriptive models as well as semi automatic annotations that were obtained by means of trained algorithms. Examples of such models are for instance those that describe addressee behavior (Jovanovic et al., 2006), or turn-taking (Padilha and Carletta, 2003a) behavior (See Figure 4.2(c)). The models of human behavior can be compared with sensor recorded behavior to determine the extent to which these models reflect reality and where opportunities for improvement exist. (see Reidsma et al. (2005b) and Bailenson et al. (2004)).

A final point we stipulate here is that virtual reality environments, such as the VMR provide the possibility to test for human observation capabilities. Poppe et al. (2007), for example, present an experiment where the VMR is used to assess the accuracy of head orientation perception from the observer in triadic situations. A few other experiments that use Virtual Environments for

elicitation and/or validation of models of (social) interaction can be found in Bailenson et al. (2001) and Slater and Steed (2001).

4.5 Machine Learning and automatic schema application

From the previous sections it turns out that automatic recognition of human behavior and events, boils down to the automatic application of annotation schemas that describe these events. If we want to investigate how this can be achieved, we enter the world of machine learning. The field of machine learning is concerned with the question how to construct computer programs that can learn from examples and that can adapt to their environment. Machine learning provides the technical basis of data mining, that is, it enables the extraction of implicit previously unknown and potentially useful information from the data. For our case, to be more precise, we want the machine learning algorithms to learn to reproduce the annotations that have been created on top of recorded sensor data and that describe phenomena related to human-human interaction in the context of meetings

A mathematical approach to the problem is the following: Imagine that there is a certain pattern, or function f that maps a certain input X, possibly consisting of more than one component, e.g. $\{x_1 x_2 ..... x_i ..... x_n\}$ onto a certain output $f(X)$. The task now is to learn what f is. Our hypothesis about f, denoted by h, is to be selected such that it approximates f as closely as possible. h is to be established based on training data T that consists of a number of examples how f maps X onto $f(X)$. This training data, as gathered in a corpus, hence should in order to be learnable, be consistent such, that sufficient similarities exist between the detectable patterns in the data and the phenomena described by the annotation schemas. The output may be a real number, in which case the process that is represented by h estimates this function. The output of h estimates the output of f. If this output is a categorical value, h has labelled X with a certain value. It has assigned it a category, or a class. The process that is embodied by h is therefore called a categorizer, or a *classifier*. The output itself is usually called a label, or a class. An example application is, for example, the recognition of hand printed characters. The input in that case is some suitable representation of the printed character, and the resulting classifier, maps this input into one of, say 26 categories.

Work in machine learning has converged from several sources and disciplines. Although, machine learning heavily overlaps with statistics and the computational properties of statistical methods are central to many machine learning algorithms, machine learning has also been inspired by artificial intelligence (i.e. parameter estimation), evolutionary models (i.e. genetic algorithms), psychological models (i.e. goal seeking by means of reinforcement learning) and even brain models (i.e. neural networks).

The goal of this section is to briefly introduce the most important terms and

aspects that are used in the area of machine learning, such that subsequent usage of terminology in the next chapters becomes familiar. However, since machine learning methods derive from so many different traditions, its terminology is rife with synonyms. What in this thesis is called input vector for example is also called, pattern vector, feature vector, sample, example, and instance. The components x_i of the input vector are throughout this thesis called features, but other names are attributes, input variables, and components.

4.5.1 Learning to Classify

We want to deduce the higher level phenomena of influence hierarchy and argumentation structure and it has become clear that we need classifiers, that learn from the annotated data how to combine those features that are able to describe the categories defined by the annotation schemas with the highest accuracy ($h = f$). Labels defined by the annotation schema are to be given to those segments in the test data that contain the most similar set of feature values for that particular label. The learning of a particular label can in turn be seen as learning a boolean function, that is, the only possible answers are '0' for wrong, and '1' for right.

For humans, the recognition of phenomena is generally a basic cognitive competence, that groups phenomena with similar features, and that groups knowledge about known exemplars of phenomena to predict aspects (e.g. behavior) for new similar phenomena (see also Section 4.2). Learning a phenomenon is a somewhat different story. In fact, there are over twelve well recognized theories that describe how humans learn (i.e. behaviorism, constructivism, neuro-science and control theory)[8]. We stay away from these and just give two examples of how human learning could happen in a reception and a selection task. In a selection task various items are presented (in ordered form). Then a positive example is shown and the learner has to select another example of which he or she assumes that it is positive. The learner receives feedback until the concept can be distinguished from the other items, based on the rule that was learned. In a reception task, each instance is shown one after another together with information whether it belongs to the searched for concept or not. After each step the learner is asked to describe the rule he or she thinks is appropriate. This process continues until the correct function, or rule is found.

For a machine there are two major categories of learning a function, or to train a classifier. Those categories are: supervised and unsupervised learning. In supervised learning one knows, similar to the reception task for humans, the output of f beforehand. Usually here all the samples in the training set T are provided to the algorithm. The goal is now to find a hypothesis h that as closely as possible agrees with f for all the instances of T. The larger T the better the h will generalize and the better its approximation of f will be. In unsupervised learning there only exists a set T and there are no associated function values provided. The challenge here is to partition T into subsets

[8]For more information about learning theories consult Thorpe and Schmuller (1954).

t_1.....t_m in an appropriate way. The value of the function f now determines the subset of T to which the input X belongs. Unsupervised learning methods have application in taxonomic problems in which it is desired to invent ways to classify data into meaningful categories. Unsupervised learning is sometimes also referred to as clustering and resembles the human reception task in a sense that both need to find out on which grounds to discern the provided examples. In our case we have created the annotation schemas that provide the output to the system. We therefore restrict ourselves to supervised learning.

4.5.2 Features and Feature selection

Features are aspects that describe phenomena and a certain combination of features can be used to differentiate between phenomena. They check a single property of the classification instances, that is the phenomena that are to be discerned. For every phenomenon that is to be distinguished from any other phenomena by a classifier, the same set of features needs to be available. Features are generally valued by either real valued numbers or discrete valued numbers.

A meeting participant for instance can be represented by the features name, age, research project, attendance rate, influence and experience. A particular participant could now be represented by a vector such as {Dennis, 30, AMI ,75, 30, 50}. If we assume that these data are collected in order to decide amongst possible employees who to let attend an upcoming meeting on multimodal corpora collection, the fact that the research project for this participant is valued AMI could be useful information. As the AMI project is renowned for its knowledge on the meeting topic, observers might therefore, based on the feature value, expect employee Dennis to be of potential value for this meeting. This way sensible decisions can be created about certain categories, such as in this case *should go* and *should not go*. If however, all of the employees are also part of the AMI project, the 'research project' feature will not be of much value anymore and as a result, another feature, for example, 'attendance rate' could become of interest. To know the appropriate set of features that is able to make the distinction that one is after is always a big challenge that is to be resolved.

Figure 4.3 gives an example of a two feature sets that can be used to make choice about who to let attend an upcoming meeting. All available employees have been labelled beforehand if their attendance is expected to be valuable (white dot), or not (black dot).

From the figure it shows that it will be very hard to make a decision on the features 'attendance rate' and 'age' and that the feature set {Influence, Experience} appears to be useful, as in this case, a rule can be created that separates the employees into the sought after categories, namely if an employee has a high of experience and is of high influence he or she is also expected to be a valuable attendant for the upcoming meeting. This knowledge, or rule that has been found is ready to be implemented in a machine. Pentland (2005) describe the following features when willing to discern profiles of meeting participants typical social behavior: speaking rate, speaking energy, speaking duration, number of participants, number of interruptions, transition probabilities between the

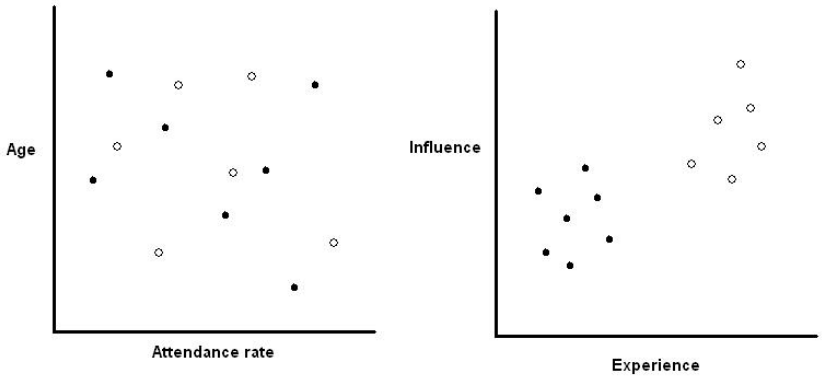


Figure 4.3: Examples of an indistinctive and a distinctive feature set

participants and the time this participant spends on holding the floor.

In a study on 'salvaging' meeting information by Moran et al. (1997), explicit marks were made in the meeting recordings to tell the person who afterwards had to create a summary where to find the information that was useful for the task. Page turnings for instance were explicitly recorded. Page turnings fulfill here the role as a feature in order to aid the activity of summarization. To automatically detect those features, that are able to help one in ones classification problem is not an easy task. Initially, to choose relevant features, one could sometimes just guess and use knowledge that is 'common sense', or available from existing literature. A different strategy is the 'wide net' strategy. This strategy tries to initially capture as many features as possible, and to find the 'relevant' features at a later moment in time. In the subsequent chapters we will use both approaches to find features that when combined can distinguish amongst the annotation categories as described in the coding schema. It is obvious that, when the deduced rules are eventually applied into systems, the fewer, and the more easily obtainable features can be used, the better.

Two important aspects to feature generation are the source and scope of the features. Eventually, all information required to generate features must come from automatic systems; however, information from annotations may be used to train systems. Also, systems are sometimes evaluated using features based on annotations, either because data from an automatic system is not available yet, or to assess the potential usefulness of a new type of feature. The scope of features depends on the application that a system will be part of. If a system runs during the meeting, only information from the past is available. In a post-processing application, all the information is available, allowing features that look forward from its current position.

Full automatic feature detection is mostly dependent on signal processing and information retrieval techniques. Typical examples can be found in the areas of computer vision and speech recognition. These techniques are applied to collected data, such as the AMI corpus.

4.5.3 Classifiers used in the next chapters

In the next chapters three different classifiers are applied in the experiments that are conducted. These are: the simple probabilistic classifier Naive Bayes, the decision tree learner J48 and from the family of generalized linear classifiers, Support Vector Machines. They are shortly introduced here.

Naive Bayes is a probabilistic classifier that uses Bayesian Formulations using prior probabilities to assign class labels (John and Langley, 1995). The underlying probability model is an independent feature model, as it assumes independency between the features.

In mathematical terms one wants to assign an instance the class label C, given its n feature values: $P(C|x_1..x_n)$. As N can be large and the feature values can be infinite one generally applies Bayes Theorem to derive a more suitable formula. After some rewriting and using the assumption of conditional independence, one can formulate the classification as follows:

$$f(x_1..x_n) = argmax_c P(C = c) \prod_{i=1}^{n} p(X_i = x_i|C = c)$$

The formula arrives at the correct classification as long as the correct class is more probable than any other class. A benefit of this approach is that class probabilities do not have to be estimated very well.

J48 is an implementation of C4.5 (Quinlan, 1993). C4.5 is a decision tree classifier that produces a decision tree consisting of non-terminal nodes and terminal nodes. The non-terminal nodes represent tests on one or more features of the data. The terminal nodes represent the outcome, or class labels. Starting from the top of the tree at each non-terminal node the classifier will test an instance for a particular feature value and push it on a branch depending on its outcome. This way the instances are divided into subgroups that adhere to the constraints of the attribute values defined by the nodes. The terminal nodes at the bottom of the tree contain the label that is assigned to the tested instance.

The creation of decision trees from the training set looks for discriminative features on which the instances can be discriminated. An example of a decision tree generating algorithm is the Iterative Dichotomiser 3, or ID3. This algorithm uses the concept of information entropy in order to determine which attribute is used for the creation of another node in the tree. Starting from the root node, the tree is expanded with the feature that has the lowest entropy in relation to the other (unused) features.

Support Vector Machines (SVM) are used for classification and regression (Boser et al., 1992). SVM's use a technique known as the 'kernel trick' to apply linear classification techniques to non-linear classification problems. Multi-class problems are now solved using pairwise classification.

For a two-class classification problem, one can visualize the operation of a linear classifier as splitting a high-dimensional input space with a hyperplane: all points on one side of the hyperplane are classified as 'yes', while the others are classified as 'no'.

Mathematically it boils down to increasing the margin **w** perpendicular to the separating hyperplane $\mathbf{w}.\mathbf{x} - b = 0$. (The offset parameter b allows to increase the margin). The support vectors are those that run parallel to the hyperplane and for which the margin is maximal (see Figure 4.4).

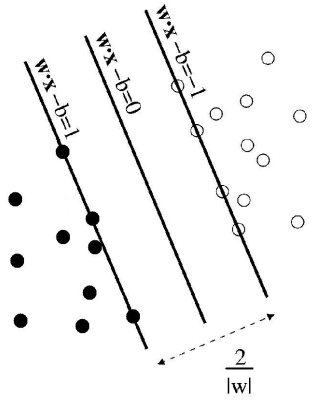


Figure 4.4: Maximizing the margins in a linear separable problem

To maximize **w**, one wants to maximize $\frac{2}{|\mathbf{w}|}$ under the constraint that the problem is linearly separable and that there will be no point between the two parallel hyper-planes. The constraints can be formulated as $c_i(\mathbf{w}.\mathbf{x}_i - b) \geq 1$, for which $1 \leq i \leq n$.. What remains is a so called quadratic programming optimization problem that in its dual form can be solved by making use of Lagrange multipliers. The SMO implementation of SVM is used throughout this Thesis.

4.5.4 Performance and Evaluation

There are several measures to evaluate the performance of a machine learning algorithm. In supervised learning the induced rule, or function h, is usually evaluated on a separate set of inputs and predetermined outputs. This set is often referred to as the testing set. h is thus said to perform, or generalize, well when it guesses well on the testing set.

A typical technique that is applied on a data set is that one repeatedly trains classifiers on different parts of the data and tests on the parts that remain. This

technique is called N-fold cross validation. If, for example, n=10, the complete data set is separated in ten sub-sets. 9 of these can be used for training, and one for testing. As a final performance result the average result is calculated from the performance on the subsets.

The output of a classifier is often reported in a confusion table, or confusion matrix. A normalized confusion matrix for a binary classification task is shown in Table 4.1.

	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

Table 4.1: A normalized confusion matrix.

Generally each column of a confusion matrix represents the instances of a predicted class, while each row represents the instances of the actual class. A confusion matrix can be used to see if a particular classifier is confusing two classes. This information is valuable as it gives directions for modifications of the used coding schema.

From the normalized confusion matrix shown in Table 4.1 one can derive a number of performance metrics. Some of them are shortly discussed here.

The metrics that will be used throughout the next chapters is the percentage of correctly classified instances on the test set, or *accuracy*. Accuracy can be defined in terms of the confusion matrix as the fraction of all instances that have been predicted correctly.

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN}$$

But there also exist other measures that can be calculated from the confusion matrix. One could be interested for instance only in the fraction of retrieved positives out of the actual collection of positives. This measure is called *recall*.

$$Recall = \frac{TP}{TP+FN}$$

A variant of Recall is Precision, that considers from all those that were predicted positive, the percentage that actually was positive.

$$Precision = \frac{TP}{TP+FP}$$

A measure that is derived from Precision and Recall is the weighted harmonic mean of both measures known as the *the F-measure*.

$$F = \frac{2.(Precision.Recall)}{(Precision+Recall)}$$

4.5.5 Feature reduction

Once a whole set of features has been collected it is interesting to know which features really matter for the classification task. To put it more formally, one might wonder what the ultimate subset Y of X is, such that the process embodied by h, does not loose (much of) its resemblance to f. Here we enter the world of feature reduction. Feature reduction mostly uses statistical techniques to reduce the number, or the dimension, of the features, while maximizing the information that is preserved in the reduced feature space. An example of one of those techniques is the leave one out method. This, rather simplistic, approach examines the contribution of each feature in terms of the performance of h, in relation to f by leaving one feature out of the total set before ranking them according to the 'error' caused by the omission. A certain threshold now defines which features are kept, and which are disposed of. A more complex method, called factor analysis takes the possibility of variable interdependency into account. Factor analysis explains the variability among observed variables in terms of fewer unobserved variables called factors. The contribution of each of the initial variables on the unobserved variables, also known as contribution to the factor, is an indication for the usefulness of the initial variable, in relation to the others that are present. The method I applied is described by Hall (1999). This method searches for features that highly correlate to the class-labels and that have a low inter feature correlation. This way the features that complement each other are preserved whereas features with a discriminative power similar to other features are being removed. (see Duda et al. (2000) for more information about the algorithms).

Chapter 5

Identification of Influential Participants

5.1 Introduction

In any initial meeting of previously unacquainted individuals who interact in the pursuit of the solution to a problem they face together, observable regularities occur. One of these is that a dominance order, or *order of influence*, is established (Rosa and Mazur, 1979; Lee and Ofsche, 1981). In psychology, dominance refers to a social control aspect of interaction. It involves the ability to influence others. One can refer to it as a personality characteristic - the predisposition to attempt to influence others - or one can use the term to describe relationships within a group. Dominance relates to the ability of *influencing* and controlling others and to power and prestige. Dominance is a hypothetical construct that is not directly observable.

A dominance order plays several important roles. It is known, for instance, that participants correlate the position in the order of the group members with the degree of influence each individual has over the group's choice of a solution. Furthermore, intelligence and judgements of high-quality contributions are generally credited to group members who rank high (see e.g. Bales (1950); Fisek and Ofsche (1970)). If, however, people become too dominant within groups, they start to exert a disproportionate influence over the group outcomes[1]. This disproportionate influence is likely to violate the regular norms in a sense that the maxims of quantity and quality (Grice, 1975) are under pressure (see also Section 2.3) and that the ideal dialogue, as described by (Habermas, 1984), where individuals possess 'the symmetrical distribution of opportunity to choose and practice speech acts' is at stake. As a result, a meeting can benefit from timely detection of this order, so that measures can be taken that allow for equal participation opportunities and that prevent further frustration of the

[1]Studies of jury decision making have for example shown that the person who talks most also has most influence over the jury verdict (McGrath, 1984)

meeting process.

This chapter investigates the extent to which it is possible to automatically extract whether some participants in four person meetings are more or less dominant than others by means of a set of easy and objectively detectable features. Throughout the chapter the terms 'rank' and 'order', as well as 'dominance' and 'influence' will be used interchangeably. The next section introduces the concept of dominance and elaborates on findings in social psychology that can steer us in our efforts to collect a useful feature set. Section 5.3 then describes the related work that we are aware of before in Section 5.4 and 5.5 two attempts are described to automatically replicate dominance hierarchies. In the first attempt we asked observers to rank meeting participants according to who they thought had most influence on the process. In the second attempt we obtained the 'ground truth' from questionnaires that were issued to the meeting participants. For both attempts the features that were used as classifier input are described. We describe how we obtained the feature values from our corpus and what the performance of various classifiers was when using the best feature combination. Section 5.6 shows the application of the resulting model in a meeting browser and in a virtual meeting environment, two applications that have clear value in the context of technological meeting assistance. The chapter finishes with the provision of some final thoughts (Section 5.7) that point out the strengths and weaknesses of the approaches that were taken.

5.2 Assessing the concept

5.2.1 Findings from Social Psychology

Social psychology has studied the concepts of dominance and influence arising from group discussions for several years. In SYMLOG (Bales and Cohen, 1979), for example, Bales distinguishes three structural dimensions in group interactions: status, attraction and goal orientation (see also Section 4.2). Goal orientation refers to the question whether people are involved with the task or rather with socio-emotional behavior, the attraction dimension refers to friendly versus unfriendly behavior and the status dimension has to do with dominant versus submissive behavior. On a checklist that he developed for observers to structure their observations in terms of these structural dimensions a number of self-report scales appeared that group members could use to rate themselves (and other group members). Eighteen, out of the twenty-six, items relate in some sense to the concept of dominance. The factors involved in these questions are meant to discern between the behaviors and can hence be regarded as features that assess concepts. An overview of the features used for the dominance category is shown in Table 5.1.

It instantly appears that most of these features are very hard to operationalize. For example to automatically determine when someone is 'purposeful' or 'alienated' is quite complex and highly dependent on human interpretative skills. For an automatic classification task, one needs easy to extract and automati-

Positive contributions	Negative contributions
active, dominant, talks a lot, extravert, outgoing, positive, purposeful, democratic task-leader, assertive, business-like, manager, authority, controlling, critical, domineering, tough-minded, powerful, provocative, egocentric, showed-off, joked around, expressive, dramatic, entertaining, sociable, smiled, warm	passive, introvert, said little, gentle, willing to accept responsibility, obedient, worked submissively, self-punishing, worked too hard, depressed, sad, resentful, rejecting, alienated, quit, withdrawn, afraid to try, doubts own ability, quietly happy just to be in group, looked up to others, appreciative

Table 5.1: Aspects of dominance according to SYMLOG

cally detectable features that can be quantified and transformed into a series of values before classification algorithms can learn something from it.

The Status Characteristics and Expectation States theory (Berger et al., 1966, 1980) was one of the first theories that considered more objectively obtainable features. This theory tried to explain how dominance orders actually are established in a group by looking at objectively obtainable identifiers of social status such as occupation, age, race, and gender. The group's measurable dominance ranking was found to be correlated with these variations in social status characteristics. Berger's theory assumed that people employ a seemingly rational strategy, that closely relates to stereotyping, based upon fixed beliefs about how the abilities associated with status and influence are distributed. Berger's perspective reinforced the view that dominance relations are more or less fixed natural arrangements that do not allow for the explication of social change.

Over the years Berger's theory was exposed to much debate and criticism. Fisek and Ofsche (1970), for instance, showed four years after Berger had made his theory known to the public, that in groups composed of participants with equal status characteristics, once a meeting is over, it will also display proportional participation differences. It was not until Lee and Ofsche (1981) showed that the knowledge of a person's social status does not *per se* need to be causal with a dominance hierarchy, before Berger could take his theory back to the drawing board. Lee et al. proved that it is not automatically said that, what might seem a self fulfilling prophecy at first sight actually comes about. Lee and Ofsche (1981) in this respect proclaimed that in typical interaction apart from social status, communication content, demeanor, and tactics of arguments are also reasonable candidates for causes of influence. The Two Process theory (Ofsche and Lee, 1981) that evolved from Lee's study points out behavioral style, or demeanor, as the core variable of interest for the assessment of a dominance ranking. Two years later Nemeth (1983) rephrased this by stating that one needs to consider the 'choreography' of verbal and nonverbal cues over time.

In the literature on dominance and influence, at least three types of nonver-

bal behavior have been identified as containing dependent variables associated with differentiated dominance rankings. These are: Proxemic behaviors, vocalic behaviors, and kinesic behaviors (see Leffler et al. (1982); Dunbar and Burgoon (2005)). We discuss them shortly.

- Proxemic behavior relates to *personal distance*. People with a higher social rank, for example, have been found to have more and better space for their use than people of a lower status (cf. Argyle and Dean (1965)).

- Vocalic behavior has to do with *participation rates*, *floor grabs*, *interruptions*, *questions*, and *laughter*. Participation rates have long been established as indicators of dominance and influence in a group (see e.g. Bales (1950)). Those that talk the most and the longest are considered more dominant than those that talk less and in shorter intervals. Willard and Strodbeck (1972) argued that to get a high status in a group one must become a high participator. To become a high participator, one should acquire the floor by being the first to speak. Rosa and Mazur (1979) put it as follows: 'To rank high in the status hierarchy, one should initiate speech often'. Another known related category in this class are interruptions. Interruptions have been called 'a device for exercising power and control in conversation' (West and Zimmerman, 1983). Those that interrupt more are hence more likely to be of higher social status. Bales et al. (1951) found that the top ranked participants also address particular other participants less and the group as a whole more than other members. The top ranked participants also receive more acts from particular others than he or she gives to them. These findings were later approved by Goetsch and McFarland (1980). Wang (2006) recently also noted that asking questions is a means to exercise power as questions allow for topic control and are an immediate allocation for turn-taking. The last category of vocalic behavior we address is laughter. For laughter it has been found that individuals of low status laugh proportionately more than people of higher status (see also Coser (1955)).

- Kinesic behavior is perhaps the richest source. It includes *facial expressions*, *eye-gaze*, *postures*, *body movements* and *gestures*. A typical difference in gestures that has been found to predict people of higher power, is for example the observation from Henley (1972), who found that people of higher power use more expressive hand gestures during speech than people of lower power. Also for eye-gaze it has been found that high power is communicated by looking more while speaking, and looking less while listening (see Dovidio and Ellyson (1985)). And last, but not least, it was Kendon (1967) who suggested that being the first to break initial eye-contact is a sign of deference, or submission.

This all leads to the expectation that there indeed appear to be detectable and objectively observable features that can predict that some people behave more dominantly than others. A tradeoff is however to be made between the

static features such as race and gender, and the dynamic features that emerge as the meeting evolves. We conclude this section with a promising quotation from Berger et al. (2002): "Several investigations suggested that the systematic elements that form the common core of the content of status stereotypes arise in some way out of behavioral inequalities that emerge in social interaction". This implies that he in fact nowadays also agrees with the findings from Lee and his colleagues.

5.2.2 Developing the schema

As stated in the previous chapter, a model is required that initially can be manually applied by humans. From these judgements the ground truth data, or class labels are deduced that, together with the feature set, form the input of the chosen classifiers.

For this exercise the model seems quite straightforward. All we are after is a ranking of the participants, such that they can be mutually compared with each other on a dominance scale. As the meetings we used were all four person meetings, it was decided that the initial model should contain four labels in the range between from '1' for most dominant and '4' for least dominant. This resulted in our initial model $m_1 = \{1, 2, 3, 4\}$. All participants in the meetings were to be assigned one of the labels, and each participant was to be assigned a different label. This was done in order to assure that a ranking should appear.

Then the question was how to assign those labels? Or what does it mean to be more dominant or influential in a meeting than someone else? Calling into mind the model of meeting aspects, that was depicted in Figure 2.1, there are at least five aspects to a meeting that can be influenced by the participants: the group, the task, the context, the process and the outcome. So participants that are influential to the task, might not be influential in the meeting process and vice versa. It was, however, decided that we abstracted away from this subdivision and that we were most interested in the dominance ranking the way people experienced the concept. It was therefore decided to initially just ask the observers to rank the participants, without mentioning the possible subdivisions and see if the observers could agree to a sufficient extent. Sufficient agreement in turn leads to more unambiguous data, and on more unambiguous data machine learning algorithms, as shown in Chapter 4, have a higher chance of successfully learning a concept.

To assure reliability of our class labels, two 'security measures' were taken. In the first place all the manual observations have been made at least four times. In our first attempt we asked five external observers and in our second attempt the four meeting participants themselves were asked to provide the judgements. By asking several people, more people will recognize themselves in the output of the resulting model, as it represents the verdict of a group of people. This can be especially useful for concepts that are hard to describe, such as dominance and emotions as the model now becomes more transferable (see also Section 4.4).

The second 'security measure' that was taken dealt with the fact that in the

average rankings from all judgements two or more people could have obtained a similar score. If this happens it shows that the observers find it hard to make a distinction between them, and as a result could better put them in the same group, that is, assign them the same label. It was decided to apply a binning algorithm that transformed all observations for a particular participant in one out of three discrete class labels: $m_2 = \{High, Normal, Low\}$ (cf. Fisek and Ofsche (1970)). The algorithm calculated the fraction of each of the individual participants in relation to all participants by dividing the sum of the valuations of all judges for each individual participant by the total amount of points the judges could spend (e.g. in the case of five participants the maximum total score is $5 * (1 + 2 + 3 + 4) = 50$), where as the minimum score for an individual participant is 5 (1 + 1 + 1 + 1) and the maximus score is 20 (5 + 5 + 5 + 5 + 5). The scores were subsequently binned into the categories High, Normal and Low by using two thresholds of 20 and 30 percent[2]. So, when a share was smaller than 20% the resulting class label was 'Low'; if the share lay between 20% and 30% the label assigned was 'Normal' and whether it was higher than 30 % the label became 'High' (see also Section 5.4.3).

A result of the binning is that in the worst case all participants could end up labelled 'Normal' and although this is nice for inter meeting comparisons, it could be a result of confused annotators that do not agree on how the schema is to be applied. To assure that this is not the case, in our initial attempt described in the next section, an extra experiment is conducted that tests if the assigned ranking values significantly differ from randomly assigned values.

5.3 Related work

Although the literature on modelling and understanding the concepts of dominance and influence in multi-party interaction seems to provide sufficient insights for the obtainment of potentially useful features, I am aware of only three attempts to automatically estimate a dominance ranking in everyday meetings.

All of these assume that (1) dominance is a high-level concept can potentially be deduced from lower level features (See Section 4.3), and that (2) these features correlate such, that models for recognition and discovery can be extracted (Basu et al., 2001; Ohsawa et al., 2002; Zhang et al., 2005).

The Influence Diffusion Model, or IDM, described in Ohsawa et al. (2002), is an unsupervised approach that generates a ranking of influential participants by counting the number of terms, reused by the next speaker from the current speaker. The model states that the person who's terms are re-used the most is the most influential. So, this model basically uses just one feature to assess the concept. In the next section, this feature is incorporated in our initial attempt to regenerate manually observed dominance rankings to determine the extent to which this feature fulfills our needs.

[2]The rationale behind these thresholds was in this case that the interval size for each of the categories was equally large. This however is a rather arbitrary choice, as other criteria, such as, for example, ending up with an equal class distribution, can be used as well

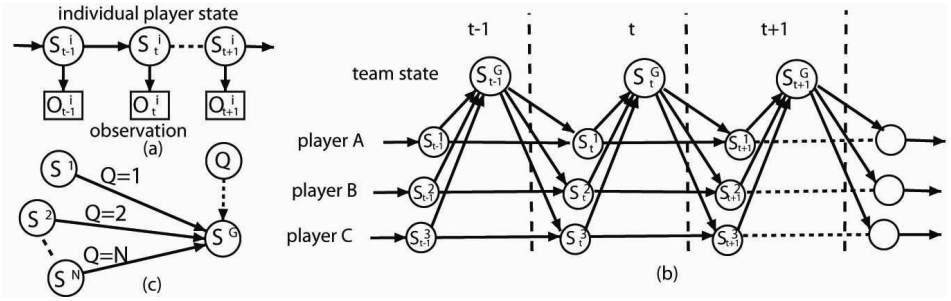


Figure 5.1: The team-player influence model (reproduced from Zhang et al. (2005)).(a) Markov Model for individual player. (b) Two-level influence model (for simplicity, the observation variables of individual Markov chains and the switching parent variable Q are omitted). (c) Switching parents. Q is called a switching parent of S^G, and $\{S^1 \cdots S^N\}$ are conditional parents of S^G. When $Q = i$, S^i is the only parent of S^G.

A second approach to automatically discover the levels of influence, also unsupervised, is described Basu et al. (2001). The features that he used were person motion energy, speech energy, voicing state, and the number of speaker turns. Basu et al. (2001) used a Dynamic Bayesian Network (DBN) as classifier that regards group interactions as a group of Markov chains, each of which influences the others' state transitions. Although this model is a tractable option, it has the limitation that it only models influence between pairs of players, and does not explicitly model the group as such.

To address this issue, Zhang et al. (2005) recently proposed a two-level influence model. This unsupervised model, called 'The team-player influence model' is a dynamic Bayesian network (DBN) with a two-level structure: the *player* level S^i and the *team* level S^G. The model is depicted in Figure 5.1. The player level represents the actions of individual players that evolve from their own Markovian dynamics (Figure 5.1 (a)). The team level represents group-level actions (the action belongs to the team as a whole, not to a particular player). In Figure 5.1 (b), the arrows up (from players to team) represent the influence of the individual actions on the group actions, and the arrows down (from team to players) represent the influence of the group actions on the individual actions.

The team state at the current time step is influenced by all the players' states at the current time step. The team state at the current time step influences the players' states at the next time step. The explicit hierarchy in the model allows for the estimation of the influence of each of the players on the team state, and the distribution of participant-to-team influence is automatically learned from data in an unsupervised fashion. This roughly works as follows: an extra hidden variable Q is added in the model to select, or switch[3], the parents for

[3]The idea of switching parent is also called Bayesian multi-nets in Bilmes (2000)

S^G by considering $\{S^1 \cdots S^N\}$ This is depicted in Figure 5.1(c). Given the observations, the probability distribution that predicts which parents are most likely to be assigned by Q can be learnt. Exactly these probabilities are of interest, as they represent the influence of a particular participant on the group.

The features used in the experiments of Zhang et al. (2005) were the number of speaker turns, the number of topic turns, and the speaking length. As 'ground truth' labels for the class values the averaged results of three observers were used. In Section 5.5 the performance of our classifiers is compared with the performance of Zhang's model, given a fixed set of features and class observations.

5.4 Attempt one: a preliminary investigation

For this study a corpus of eight four-person meetings was used.[4]. The meetings varied in length between 5 and 35 minutes. A total of 95 minutes were collected. Different kinds of meetings were used, including group discussions where topics had to be debated, discussions about the design of a remote control, book club meetings and PhD. evaluation sessions.

5.4.1 Testing the annotation schema

Ten observers were asked to rank the participants of the meetings with respect to their perceived dominance. Each person ranked half of all the meetings. This resulted in a total of five rankings for every meeting. Observers were told to rate the four people involved in the meeting on a dominance scale that ranges from 1 to 4 and that each mark had to be assigned once. The observers received no information about what was meant with the term 'dominance'. This, in order to assure that the widest notions of the concept were embedded in the judgements. The results are shown in Table 5.2.

The first cell shows that in the first meeting (M1), judge A1 thought that the most dominant person was the one corresponding to the fourth position in this list, second was the first person in this list, third the second person in the list and least dominant was the third person in the list: 2,3,4,1. If one looks at the judgements from the other observers for this meeting (A2 to A5), by comparing the different columns for this first row, one can see that A3's judgments are identical to A1's. All but A4 agree that the fourth person on the list was most dominant. All but A5 agree that the third person was least dominant. All but A2 agree that the first person was the second dominant person. This seems to suggest that on the whole judgements were largely consistent across judges at first sight.

[4]The first three meetings are meetings from the AMI project, M1 and M2 are the AMI pilot meetings AMI-Pilot-2 and AMI-Pilot-4, M3 is a meeting from the AMI spokes corpus (AMI-FOB 6). The last five are meetings recorded for the M4 project (cf. http://www.m4project.org: M4TRN1, M4TRN2, M4TRN6, M4TRN7 and M4TRN12)

	A1	A2	A3	A4	A5	'Average'	'Variance'
M1	2,3,4,1	3,2,4,1	2,3,4,1	2,1,4,3	2,4,3,1	2,3,4,1	8
M2	2,3,4,1	2,3,4,1	2,3,4,1	2,3,1,4	3,2,4,1	2,3,4,1	8
M3	2,1,3,4	3,1,2,4	2,1,4,3	3,1,2,4	1,2,3,4	2,1,3,4	8
M4	2,4,3,1	2,4,3,1	1,4,2,3	2,3,4,1	1,4,3,2	1,4,3,1	4
	A6	A7	A8	A9	A10	'Average'	'Variance'
M5	4,3,2,1	4,3,1,2	3,4,1,2	4,3,1,2	3,4,1,2	4,3,1,2	6
M6	1,3,2,4	1,4,3,2	3,1,4,2	3,1,4,2	1,3,4,2	1,3,4,2	12
M7	1,4,3,2	2,4,3,1	3,2,1,4	2,4,1,3	1,4,3,2	1,4,2,3	14
M8	1,2,4,3	1,4,2,3	2,1,3,4	2,1,3,4	1,2,4,3	1,2,3,4	12

Table 5.2: Rating of meeting participants for all the annotators per meeting.

To examine this more closely we compared the variance of the judgements with the variance of random rankings. If the variance of the annotators is smaller than the variance of the random rankings, we have a strong indication that people agree on how to create a dominance ranking (cf. Zhang et al. (2005)).

The initial step to calculate variance is to calculate the mean, or the average. The average rankings were assessed by first summing up the scores for each of the participants and then re-ranking them. This results for the first meeting in scores 11, 13, 19 and 7, with translates in an overall ranking of 2, 3, 4, 1. In case of similar scores, we scored them an equal rank by giving them both the highest value. The next one highest in the ranking was ranked with a gap of two. Example: if the sum of the total scores ended up 8, 10, 12, 10 the resulting ranking became 1,2,4,2.

As a measure for the variance the sum of all the (absolute) differences of each of the observers (A^i) with the corresponding average was calculated. The difference with the average was calculated as the sum of the pairwise absolute differences for all the annotator values of the meeting participants A_p with their corresponding average value $Average_p$. See Table 5.2 for the results.

$$\text{'}Variance\text{'} = \sum_{i=1}^{5} \sum_{p=1}^{4} |A_p^i - Average_p|$$

In this case A1 and A3 judgments are identical to the average. A2 made different judgments for the first person (scoring him as 3 instead of 2) and the second person (scoring him as 2 instead of 3). So this results in a variance of 2 adding up the variance 4 and 2 of judges A4 and A5 respectively this ends up in an overall variance of 8 for judgements on the first meeting.

When comparing the variance of the judges with the variance resulting from randomly generated rankings, the distribution of the variance of the annotators ($\mu = 9$, $\sigma = 3.38$, $n = 8$) lies far left of the distribution coming from randomly generated rankings. ($\mu = 17.72$, $\sigma = 3.60$, $n = 1000$). Although the sampling size is relatively small, a 1-sided T-test shows that the two distributions differ significantly ($F(7.12) = -7.26, p < 0.001$). From this it was concluded that

the observers agreed sufficiently on dominance rankings, such that automatic replication could be worth the try.

5.4.2 Collection of Features

We assumed that dominance can be regarded as a higher level concept that might be deduced automatically from a subset of lower level observations, similar to the assignment of the value for dominance by humans on the basis of the perception and interpretation of certain observed regularities.

After the observers that rated our corpus had finished their ratings, we asked them to write down a list of at least five aspects which they thought they had based their rankings on. The following features were mentioned.

> Dominant is the person: who speaks for the longest time, who speaks the most, who is addressed the most, who interrupts the others the most, who grabs the floor the most, who asks the most questions, who speaks the loudest, whose posture is dominant, who has the biggest impact on the discussion, who appears to be most certain of himself, who shows charisma, who seems most confident.

From the features identified by the observers we are again confronted with the fact that certain features, such as *charisma* and *confidence* are very hard to measure and to operationalize. Most of this is due to the fact that a proper scale does not exist. Some other features, however, do appear to be suitable for our task.

For this initial exploration, a trade-off was made between the available time to conduct the annotations, the features that were identified in the literature, as described in Section 5.2.1, and those that were pointed out by the observers. Deliberately no semantically oriented features were used. The following easily obtainable features that possibly could tell us something about the dominance of a person in relation to other persons in meetings were collected:

A *floorgrab* was defined each time a participant started speaking after a silence larger than 1.5 seconds. A *successful interruption* was counted if a speaker A starts talking while another speaker B is talking and where eventually speaker B finishes his turn before speaker A. This procedure, to calculate an interruption, has, for example, been described in Leffler et al. (1982). No distinction is made between overlap and interruption.

Most of the features appear as simple metrics with variations that measure the amount to which someone is involved in the conversation and how others allow him/her to be involved.

5.4.3 Data Acquisition and Preprocessing

For each of the eight meetings that were ranked by our observers, we collected the values for the measures identified in the previous section. This was done on the basis of simple calculations on manual annotations and on the results of

The speaking time in seconds (STS) The number of turns in a meeting (NOT) The number of words spoken in the whole meeting (NOW)
The number of successful interruptions (NSI) The number of times interrupted (NTI) The ratio of NSI/NTI (TIR)
The number of times the person grabbed the floor (NOF)
The number of questions asked (NQA)
The number of times addressed (NTA) The number of times privately addressed (NPA)
The person's influence diffusion (IDM) Normalised IDM by the amount of words spoken. (NIDF)

Table 5.3: Features that were manually collected for the initial attempt to automatically assess a dominance hierarchy.

some scripts processing the meeting transcriptions[5]. With respect to addressee annotation 25% of the data was not annotated due to the cost involved[6].

To make the feature values for the same features inter-meeting comparable an approach was followed similar to that for the class values (see Section 5.2.2). First the feature values were corrected for the meeting length by computing their fraction in relation to all the values for that feature in a meeting. Then the resulting fractions were again binned in three different category bins : 'High' ($F^{'}_{Pn} > 35\%$), 'Normal' ($15\% < F^{'}_{Pn} < 35\%$), and 'Low' ($F^{'}_{Pn} < 15\%$).

Table 5.4 shows the value of the NOW feature ('The number of words used' per participant per meeting) before and after applying the process. If we look at the number of words used for participant 2 (P2) and participant 4 (P4) in Meeting 1, we see that they both end up labelled as 'High', although they did not speak the same number of words. They, however, both used much more than 90000 words, which is much in comparison with P1 (38914) and P3 (26310), both ending up classified as 'Low'.

The class labels that determine the dominance level were obtained by applying the binning process described in Section 5.2.2 on the averaged observed dominance level depicted in Table 5.2. This resulted in a data set of 32 samples with twelve samples receiving the class label 'High', ten 'Normal' and ten 'Low'. The share of the most frequent class label ('High') was used as a baseline for our classification results. If our classifiers could outclass the baseline of in this case, 37.5%, it would prove that we could, better than randomly, automatically predict the dominance values for the participants.

[5] All transcriptions used were created using the official AMI and M4 transcription guidelines of those meetings (Moore et al., 2005; Edwards, 2001)

[6] Addressee information takes over 15 times real time to annotate (Jovanovic et al., 2005)

	NOW before preprocessing				NOW after preprocessing			
	P1	P2	P3	P4	P1	P2	P3	P4
M1	38914	93716	26310	98612	low	high	low	high
M2	33458	11602	14556	37986	high	low	low	high
M3	3496	7202	8732	2774	low	high	high	low
M4	2240	1956	4286	7642	low	low	normal	high
M5	4470	1126	9148	1974	normal	low	high	low
M6	2046	17476	1828	4058	low	high	low	high
M7	4296	6812	8258	1318	normal	high	high	low
M8	1586	13750	1786	1540	low	high	low	low

Table 5.4: The feature 'Number of Words' before and after preprocessing for person 1,2,3 and 4 respectively for each meeting.

5.4.4 Results

We wanted to predict the dominance level of the meeting participants, in accordance with Occam's razor (Blumer et al., 1987), by trying to explain as much as possible with as little as possible. The fewer, and the more easily obtainable, the features that are required, the easier it would be to eventually provide all information to a system that has to decide upon the eventual class label. This way the risk of over-fitting our model to the data is reduced as well. The risk of over-fitting is here, due to the small amount of samples, and the relatively large feature size, quite high. To decrease the number of feature reduction algorithm described in Hall (1999) was applied. (See also Section 4.5.5)

The two most discriminative features appeared to be NOF and NOT. Table 5.5 shows the results from each of the three classifiers on these two features. All the results are obtained using ten-fold cross validation.

Classifier	Accuracy
NB	75.0%
J48	68.75%
SVM	75.0%

Table 5.5: The results of the three classifiers on our initial attempt to automatically assess the dominance levels of individual meeting participants

The obtained performance of 75% is much higher than our 37.5% baseline. This theoretically would mean that, given the number of times the meeting participants grab the floor after a silence together with the number of turns a participant has, two of the three classifiers are in 75 % of the cases able to correctly classify the behavior of the participants as being 'Low', 'Normal' or 'High' on dominance. Due to the small number of samples one, however, has to be very careful about generalizing the results.

To become a little more certain, the 90% confidence interval was computed from the individual folds. This confidence interval showed that in 90% of the cases the performance lies between 62% and 88%. With a lower bound of 62.5%, this performance is much higher than the 37.5% baseline. This confirms the expectations that we can, to a certain extent, automatically assess a dominance hierarchy with relatively few and easily obtainable features. The fact that we would over fit our classifier when using all the features indeed appeared when we trained on all the features. Tenfold cross validation resulted in that case in a performance of 50%.

The confusion matrix from the 75% correctly predicting SVM classifier is shown in Table 5.6.

	Low	Normal	High
Low	8	1	1
Normal	2	7	1
High	0	3	9

Table 5.6: Confusion matrix using the features NOF and NOT. The rows show the actual labels and the columns the labels resulting from the classifier.

From the confusion matrix it can be seen that the classifier performs better on the classes 'Low' and 'High' than on the class 'Normal'. This seems in line with the intuition that people showing more extreme behavior are easier to classify.

The next section describes our second and more elaborate attempt to automatically assess the participants' influence on the meeting process.

5.5 Attempt two: Expanding Feature and Data set

This section describes the second attempt to automatically assess dominance rankings of the meeting participants. In this attempt a larger meeting corpus is used as the basis for experiments, and the obtained results from our 'static' classifiers are compared with the results obtained from Zhang's 'dynamic' model.

5.5.1 Collection of Features

The features that were used include a selection of the features described in the previous section, as well as some that were used in the work of Zhang et al. (2005), and some newly designed ones. The features in the used feature set relate both to the demeanor of the participants and to the status of the participants. They can be grouped into three categories: individual speech behavior, interaction behavior, and semantic-based features. An overview of all the features that were used is shown in Table 5.7.

The number of turns in a meeting (NOT)*
The average number of words per turn (NWT)
The average duration per turn in ms. (ADT)
The number of times the person grabbed the floor (NOF)*
The number of successful interruptions (NSI)*
The number of times interrupted (NTI)*
The predefined role of the participant (TPR)
The number of initialized topics (TNT)*

Table 5.7: Features that were collected for the second attempt to automatically assess a dominance hierarchy.

Two new features, the more semantically oriented features, were used for this trial. The predefined participant roles were obtained from the AMI meeting metadata. As already described in Section 1.3.1, there are four types of roles assigned to the meeting participants in the AMI Corpus: Project Manager, Industrial Designer, User Interface Designer and Marketing Expert. A turn was again defined by a complete utterance without silences longer than 1.5 seconds and containing at least one word. A successful interruption was also defined similar to the previous attempt, except for the fact that in addition the turn of speaker A now had to be at least three words long, rather than one. Most of these features were again obtained from the manual transcriptions of the meetings.

For the TNT feature the numbers of topics initiated by each of the participants that were resumed by another participant were calculated. The topics were obtained using probabilistic latent semantic analysis (PLSA) in a similar fashion as described in Zhang et al. (2005). PLSA is a language model that projects documents in the high-dimensional bag-of-words space into a topic-based space of lower dimension. Each dimension in this new space represents a 'topic', and each document is represented as a mixture of topics. In our case, a topic corresponds to one speech utterance. Therefore, the topic boundary is equivalent to the utterance boundary. PLSA is thus used as a feature extractor that could potentially capture 'topic turns' in meetings[7].

Two versions of feature sets were created: one version where all the feature values were normalized, and one where all the feature values were normalized and binned depending on their fraction in relation to the other participants. The features that had to be normalized in order to make them inter-meeting and inter-person comparable are indicated by '*'. The procedures that were carried out are further similar to those described in the previous section.

For Zhang's dynamic model, the features were manipulated as follows. For the first feature, NOT, the feature sequence consists of binary values, one if the person speaks and zero otherwise. The NOF, NSI and NTI features were treated similarly. Figure 5.2 (a) illustrates this. For the NWT and ADT feature, the

[7] see Hofman (2001) for more information

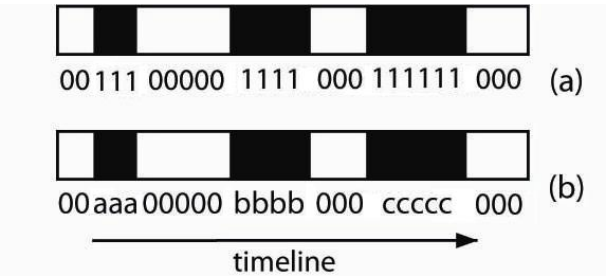


Figure 5.2: Illustration of the sequential features which serve as input to the dynamic model: (a) a sequence of binary features. For example, one indicates *speaking*, and zero indicates *silent*. (b) A sequence of number of words (or utterance duration) features, where a, b, c indicate the number of words (or the speaking duration) in one utterance separately. We repeat the same value within the same utterance. The value for the silence segments was set to zero.

value was repeated within the same utterance. The number of words for the silence segments was set to zero (Figure 5.2(b)). Note that this feature representation is effectively using non-causal information. Finally, the *role* feature was not used for the dynamic model, as its value is constant for each participant over the entire meeting.

5.5.2 Data Acquisition and Preprocessing

For this attempt, all meetings used were project scenario meetings from AMI meeting corpus, recorded at TNO-Soesterberg in the Netherlands. The set comprised 40 meetings of 10 different design teams with an average duration of 30 minutes each. Figure 5.3 shows the view from one of the overview cameras for a typical meeting.

Figure 5.3: A view from an overview camera of a typical meeting recorded at TNO-Soesterberg.

To obtain the class label ground-truth, both for evaluation and for training of supervised methods, we used the rankings of the individual participants provided by questionnaires. For all the meetings, questionnaires were filled in by the participants on which a number of questions had to be completed. One of the questions asked participants to rank all of the meeting's participants, including themselves, from most to least influential by assigning them unique nominal values ranging from one (most influential) to four (least influential). Participants were not allowed to rank people equivalently. This is in contrast to the attempt described in the previous section where external observers were used. It also contrasts with the earlier approach from Zhang et al. (2005), as he assigned real numbered influence values to participants in his earlier work. The collected permutations of the numbers one, two, three and four, were again quantized into three classes as described in Section 5.2.2. The resulting data set has a total of 160 labels (40 meetings times four participants) resulting in 34 observations for 'Low', 91 for 'Normal', and 35 for 'High'.

5.5.3 Results

The static classifiers are, in contrast to Zhang's dynamic model, equipped for post-meeting processing. This implies that they use features of which the values are summed and normalized when the whole meeting, or a meeting segment, is over. The static classifier models are hence unable to dynamically 'track' the influence online.

Results for the static models

The static classifiers that were used were those described in Section 4.5.3: Support Vector Machines, J48, and Naive Bayes. As the problem was modelled

as a classification problem, the percentage of the correctly classified instances (accuracy) and the confusion matrix are mentioned for both models. Tenfold cross-validation was in every case applied while determining the results. The accuracy figures that were obtained with the data are shown in Table 5.8.

	Original data	
Classifier	Normalized (%)	Normalized/Binned (%)
J48	46.25	57.50
Naive Bayes	66.25	61.25
SVM	61.25	60.00

Table 5.8: Results of the static model on the original 'unbalanced' data set.

It appears that the static models are performing quite badly on the unbalanced training set. Given a baseline of 57% (91 out of the 160 observations were labelled 'Normal'), the results are far from good.

As the skewness of the label distribution is a result from the binning algorithm, a set of 100 different balanced versions was created to test the sensitivity for unbalanced training. (More than one balanced version was created to preserve the distribution of the feature values.) The resulting data set contained 34 observations for all of the class labels (102 in total). Table 5.9 shows the averaged performances including standard deviations for the balanced data sets.

	Balanced	
Classifier	Normalized (%)	Normalized/Binned (%)
J48	52.18 (5.14)	54.93 (5.13)
Naive Bayes	59.65 (2.98)	60.16 (3.40)
SVM	58.78 (3.64)	58.45 (3.85)

Table 5.9: Results of the static models on balanced data sets (standard deviation in brackets).

As we now have a baseline of 33% due to our balanced training sets, it appears that the results as shown in Table 5.9 are much better than those in Table 5.8.

To gain a little more insight into the results, the best performing (normalized and binned) feature set for each of the classifiers were examined for the performance for the individual features. The results are summarized in Table 5.10.

From Table 5.10 it follows that, in particular, the feature NOF by itself is unable to outperform the naive baseline of 33%. The TNT feature on the other hand seems, together with the NOT feature to be quite robust and useful. This can be seen in the table by looking at the highest value in the second column, in combination with the lowest value in the third column.

	Alone (%)	All Except (%)
NOT	57.84 (NB)	61.76(NB)
NWT	50.98 (SVM)	65.69(NB)
ADT	56.86 (NB)	70.59(NB)
NOF	31.37 (SVM)	69.61(NB)
NSI	50.00 (J48)	70.58(NB)
NTI	42.16 (J48)	68.63(NB)
TPR	46.08 (NB)	64.71(NB)
TNT	57.84 (NB)	57.84(NB)
All features	**70.59**(NB)	

Table 5.10: Performance of individual features for the balanced set (normalized and binned) with the best performance (70.59%) and the performance for all features except the feature mentioned in the row (Classifier in brackets).

Post hoc feature subset evaluation (by applying the algorithm described in Hall (1999)), revealed a best subset containing the features NOT, ADT, TPR and TNT. Using only the resulting subset, a best performance of 69.61% was achieved using NB (not shown in Table 5.10). We conclude the results on the static model by presenting the confusion matrix for our best result (70.59%), which used all features, in Table 5.11.

→ Classified as	Low	Normal	High
Low	23	7	4
Normal	8	23	3
High	0	8	26

Table 5.11: Confusion Matrix on our best run (=70.59%) using Naive Bayes.

From the confusion matrix it follows that 'Low' influence persons are sometimes labeled as 'High', whereas 'High' influence persons are never labeled as 'Low' influence.

Results for Zhang's model

The dynamic model was trained with all (except the 'role') features individually. All features were extracted at 5 frames per second. For example, for a 5-minute meeting, the total number of feature frames is 1500. The learned influence value that results from the dynamic model (α_i) is a real value in the range between 0 and 1. This value was transformed into one of the three discrete class labels using two thresholds (th_1, th_2). For all experiments ten-fold cross-validation was used. All cases were tested with different parameter configurations (*i.e.* th_1, th_2). Reported in Table 5.12 are the mean accuracy and the standard deviation.

Method		Accuracy (%)
Individual Features	NOT	56.25 (4.23)
	NWT	57.50 (5.27)
	ADT	61.25 (4.84)
	NOF	48.75 (4.77)
	NSI	48.13 (4.16)
	NTI	45.00 (3.44)
	TNT	53.50 (4.54)
Fusion	average	54.38 (3.94)

Table 5.12: Results obtained with Zhang's model on different features (standard deviation in brackets).

For feature fusion, a naive averaging method was used: $\alpha = \frac{1}{K}\sum_{i=1}^{K}\alpha_i$, where K is the number of features.

5.5.4 Reflection on the results

After looking at the outcomes of Zhang's model (Table 5.12) in comparison to those obtained by the static classifiers (Table 5.10), at least the following issues are worth commenting upon: it appears that the best combined feature performance of the static model (70.59%), outperforms the best performance of the dynamic model (54.38%). The best individual feature for the dynamic model turns out to be ADT, while the best individual features for the static model seem to be the NOT and the TNT. This indicates that the best feature using a dynamic model is not necessarily the best feature using static models. For each individual feature, it is hard to say which classifier is better. For example, the performance of the NOT feature whilst using the dynamic model is better than using SVM, but worse than using NB. The best subset, containing just four out of the eight examined features, resulted for the static classifiers in a performance nearly equal to that for the complete feature set. With respect to the amount of effort one wants to invest on feature extraction, this is certainly something to take into account. With respect to the significance of the results, I would again like to mention that although our sample size is considerably larger than in the attempt described in the previous section, it is still relatively small when compared to a typical classification problem.

On a general level the differences between the dynamic model used by Zhang and the static model lie in the fact that the static model is comparatively quite fast and that it is able to combine several features and requires feature values calculated over the whole meeting. And although the dynamic model, on the other hand, can deal with dynamic feature value updates, it cannot output the influence of each meeting participant at any moment of the meeting, while, as described in the subsequent section, the static models can do this while applying some heuristics.

5.6 Application

Typical applications of systems that track the influence levels of participants are other systems that use the influence information in order to inform the meeting participants or a meeting chairman about this. With this information a chairman could alter his style of leadership in order to increase the meeting's productivity. Combined with other information, recommender systems could be created that directly suggest how to change the leadership style. One could even think of a virtual chairman which is able to lead a meeting all by itself, maintains a good balance, gives turns, keeps track of a time-line and most importantly: keeps the meeting as pleasant, effective and efficient as possible.

Also, the direct reporting of the deduced information to the participants themselves could prove useful. Pentland (2005), for instance, reports the usage of displays that reflect dyadic relationships to show insights into the 'role' played by the participants. In work from DiMicco (2004) a system called Second Messenger is described that shows real-time text summaries of participants' contributions. It appeared that after increasing the visibility of the less frequently speaking group members, these started to speak more frequently than before, whereas the more talkative people started to speak 15% less.

Inspired by the obtained results described in the previous two sections, a very simplistic model was crafted based on three very easily 'live' detectable features: NOF, NOT and NSI. The model grants one point for each turn a participant takes in a meeting and if the turn is acquired, either after a silence greater than 1.5 seconds, or by an interruption, another extra point is given. In this way the model allows for 'live' tracking of the influence levels. This section presents two applications of the developed model: a JFerret meeting browser (see Section 3.4.2) implementation and an implementation in the Virtual Meeting Room (VMR) (see Section 4.4.3).

5.6.1 JFerret implementation

A first implementation has been created for the JFerret meeting browser, developed by Wellner et al. (2004), which enables people to access meeting information. Here the influence levels are shown over the meeting depicted by a graph (see Figure 5.4).

All the obtained points for the participants of a particular meeting are counted in an adjustable time window. This provides the opportunity to view the output either as a set of cumulatively increasing lines (whole meeting period), or as a set of lines revealing more about the change of the output over time, such as a five-minute time period, as shown in Figure 5.4. It should be noted that, due to the fact that a trained classifier is used, once the meeting is over, the resulting heights of the participants' levels indeed correspond to the averaged observed value (Meeting M3 in Table 5.2).

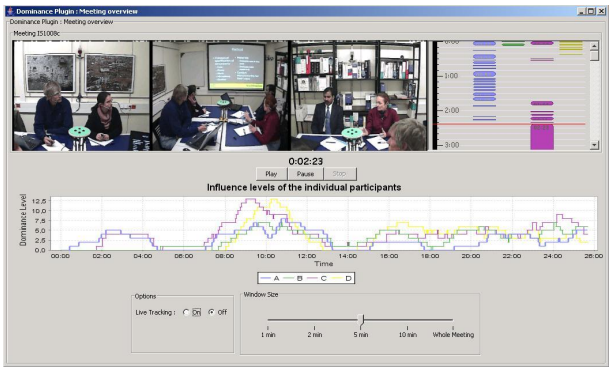


Figure 5.4: A graphical visualization of the calculated influence levels.

One could envision that, one day, if the browser is used by managers interested in the performance of their employees, the influence levels plug-in could provide valuable information. If just and valid arguments were put forward on the one hand and the person was not influential in the given setting on the other hand, this might also be a point to address. Also as a preparation task, looking over the behavior of influential participants in a previous meeting might prove useful when selecting someone to attend an upcoming meeting with these same participants.

5.6.2 VMR Integration

Another implementation has been realized in the Virtual Meeting Room (VMR), a copy of the smart meeting room at IDIAP, developed at Twente (Reidsma et al., 2005b). The VMR was developed for schema validation, signal replay, as a remote conferencing application, and to serve as a test environment for software agents. This virtual meeting room can be easily augmented with the relative influence levels, as in this case depicted in Figure 5.5 by the size of the black balls shown in front of the participants. This in addition to, for example, the domes surrounding the participants' heads that provide information about their gaze behavior, see Nijholt et al. (2006).

5.7 Final Thoughts

In the second attempt it appeared not possible not reproduce the results from the initial attempt (70.59% vs. 75%). Reasons for this could be numerous. It could, for example, be that the meaning behind the obtained class labels differed due to the fact that in the initial case they were provided by external observers and in the second case by the participants themselves. Although there is no evidence that the 'subjective' route differs from an 'objective' one, the initial cost involved could be a reason to take a more 'subjective' route, as

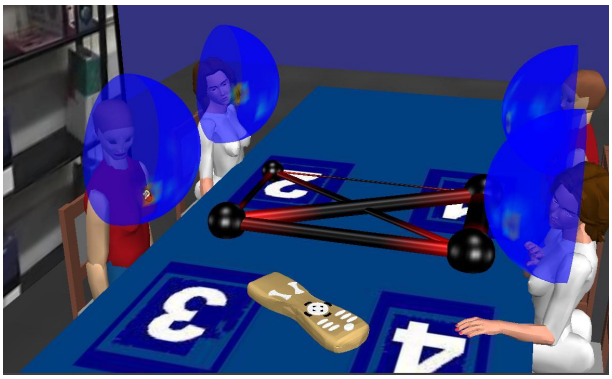


Figure 5.5: A visualization of the calculated influence levels in a Virtual Meeting Room

it is a costly enterprize to have all meetings viewed and annotated (preferably more than once). Another reason from a different category could be that, more, different, and longer lasting meetings have been used, resulting in a significantly larger number of data samples (102 vs. 32 for the static model). The more data samples, the more different examples are incorporated in the training set and the better the training set mirrors the real world. A training set that is too small, might not contain some harder to detect instances.

On a more general level it is important to consider what precisely has been measured when the observers were asked to rank the participants on a dominance scale[8]. Do the observations of the observers resemble the observations from the participants themselves? Did the observers all have the same notion of the concept, or did some put more focus on the task, whilst others focussed more on the process? Even more hypothetical, are the answers to the questions in the questionnaires independent on the time interval between the end of the meeting and the moment of filling in. It appeared that irrespective of these issues, the rankings could be quite successfully reproduced.

When looking at the rankings, in the first attempt these could be predicted quite well with the usage of the NOF feature, whereas in the second attempt this was totally the opposite. And although this could be because in the second attempt the NOF feature was tested alone, where as in the first attempt the performance was calculated in combination with the NOT feature, this hypothetically could also be proof of the fact that our class labels represent different notions.

Our approach to the problem has been to find an appropriate *set* of features that together realize the successful recognition of the sought after class labels. The features we used were distilled from our earlier work combined with the social psychological literature we studied, and intuitively all seemed appropriate.

[8]See Butt and Fiske (1968) for an interesting expose.

A broader spectrum of features, possibly from other modalities such as vision, might eventually lead to better results. Facial expressions, from which eye-brow positions, and smiles could be distilled (Mignault and Chaudhuri, 2003) could prove beneficial, as well as more semantically oriented features, such as 'Who is using the strongest language?', or 'Who gets most suggestions accepted?'. Although these features seem logical and might increase the performance indeed, one does have to realize that to be able to measure these, costly and complex inferencing systems have to be developed. A trade-off exists here between the amount of time one wants to invest in order to collect all these features and the potential benefit that they bring. Especially, if one cannot predict beforehand how good a particular feature will be, not on its own, nor in combination with other features. I have tried to limit the feature set to those features that were straightforward, easy to obtain, whilst maintaining a maximal variety. A positivist approach was taken by including all those features that, at least in one paper, have proven to be positively correlated to the concept of dominance. Post-hoc feature examination in turn, as shown in Section 5.5.3 and Table 5.10 provides the conclusive information about which subset of collected features to go for in any resulting application.

When considering the class labels we used, it was already mentioned that a drawback of the binning algorithm is that it could result in an unequal distribution of the labels. For the dynamic model we found, for example, that using different thresholds yields better results. Hence, in future experiments one could decide to modify these in order to end up with an equally balanced corpus. We did not do this, as we wanted our results, obtained in the second attempt, to be comparable with the results in our initial attempt. Another thing that could be worth considering is to leave out those observations where the observers strongly disagreed, as this could be beneficial for our training set. A drawback of this would be that the sparsity of the samples will increase even further.

Another aspect is the performance measure we used. We looked at exact prediction of the correct class labels, whereas from an end user point of view, the interpersonal findings (Was A more influential than B?) might be of greater importance. Although we are certainly satisfied with the results achieved, it is clear that many choices can be made along the way.

The next chapter will show another approach to automatically assess a higher level concept in the meeting domain, namely that of argument structures that evolve in meeting discussions.

Chapter 6

Revealing Argument Structures

6.1 Introduction

Argumentation has been regarded as the primary means of making progress by human beings (van Gelder, 2002). It is pervasive in everyday life and plays an important role in human communication. Argumentation research spans from argumentation found in research papers to knowledge representation tools supporting the construction of rhetorical arguments, where a general aim for argumentation theory in general is to make people think critically about the arguments of others, and to create a better, more measured argument of their own.

Argumentation in the view of Perelman and Olbrechts-Tyteca (1969) comes into play in disputes where values play a part in order to reach agreement. These disputes can neither be resolved by empirical criteria (or sensory proof) nor by logical criteria (formal proof). Sometimes, the word argument is used to mean dispute, or fight as in the sentence 'The parents got into so many arguments over the child's problems that they finally got a divorce.' In our point of view argumentation is about rational persuasion, or to put it as Walton (1996) did: we see an argument as a reasoned attempt to justify a conclusion.

According to Pallotta et al. (2004) a meeting can be seen as a multi-party decision making process: a collaborative problem solving process, where people follow a series of communicative actions in order to establish common ground. Much argumentative discourse usually takes place in collaborative problem solving processes such as design meetings (cf. Buckingham Shum (2003); Pallotta et al. (2006)). Most of the argumentation in meetings and everyday conversations occurs when there is a potential conflict of opinion or attitude between participants. This conflict, generally embodied in a discussion, originates from different judgements of alternatives. For a discussion to emerge one feels obliged to produce a justification of ones beliefs, attitudes or actions; there has to be

a ground for questioning. A situation that provides an occasion to challenge somebody's statements, that is, the situation gives rise to a doubt.

This section describes an approach that is able to capture the decisions of a meeting as well as the lines of deliberated arguments in an automatic fashion. The intention is not to formulate an opinion about the *contents* of the argumentation, but rather to identify the relations and the forthcoming *structure* between the arguments. We are therefore neither interested in features that say something of the quality of argumentation[1], but rather in features that can distinguish the more objective and structural aspects. With the resulting annotations systems one should be able to find answers to questions that relate to the decision making process. These are questions such as: *Who was in favor of the proposal from X? Were there any objections raised to the final conclusion?*, or, *Were there any other solutions debated?*.

Section 6.2 describes the investigated theories that inspired the creation of the resulting Twente Argument Schema (TAS). TAS is elaborately described in Section 6.3.1. Its constituents are charted, the reliability of the schema is investigated and the resulting corpus that was created by means of manual application is described. Section 6.4 then describes related work that has been conducted along three main steps that eventually lead towards the automatic application of the TAS schema. These are: the detection of discussions in a meeting, the segmentation of the uttered speech into chunks that can be labelled, the detection of the relations between the identified utterances, and finally, the labelling of the detected relations. My efforts confined themselves to two of these steps. Section 6.5 elaborates on the effort that was put into the automatic application of TAS unit labels to the predefined individual speech utterances and Section 6.6 shows the attempts that were taken to label predefined relations between these utterances with TAS relation labels. Possibilities for applications are described in Section 6.7. An implementation is shown where the resulting argument diagrams are embedded as a plug-in in a meeting browser. A small usability study is also reported on that evaluated the usability of these diagrams in the context of seeking answers to questions. The last section, Section 6.8, presents, similar to the final section in the previous chapter some final considerations about the procedure followed and discusses some benefits, as well as some drawbacks of the approach that was taken.

6.2 Structuring Argumentation

The simplest argumentation consists of just one argument, but the structure of argumentation can also be much more complex. The argumentation structure of a text, speech or discussion is determined by the way in which the reasons advanced hang together and jointly support the standpoint that is defended (Van Eemeren, 2003). To model sequential processes, that 'hang together' in an orderly way, graphs, structure diagrams and flow charts are widely used

[1] Burroughs et al. (1973), for instance, shows that when people receive more eye gaze from naive group members when they provide high quality arguments.

in all kinds of scientific domains. These methods provide a 'broad picture' of the 'structure' of the events or phenomena that one wants to investigate or communicate.

The primary tool currently in use to give an account of argument structure is the argument diagram. An argument diagram is generally a graph containing a set of points or vertices joined by lines or arcs. The points (nodes) are used to represent statements and conclusions of the argument, the lines (arrows) join the points to represent steps of inference. An argument diagram provides a map or snapshot of the overall flow and structure of the extended chain of reasoning in a given passage of discourse containing argumentation. A typical argument diagram is a map of the overall structure of an extended argument. It charts evidence that can be reproducibly verified as being present or absent in a given case. The diagrams hence provide a point of view or judgement that anyone is free to accept or reject. They often serve as a basis for criticism and reflection of the discussion, but they can also be used for various other purposes. Some of them are mentioned below:

- Argument diagrams provide a representation leading to quicker cognitive comprehension, deeper understanding and enhances detection of weaknesses (Schum and Martin, 1982; Kanselaar et al., 2003).
- Argument diagrams aid the decision making process, as an interface for communication to maintain focus, prevent redundant information and to save time. (Yoshimi, 2004; Veerman, 2000).
- Argument diagrams keep record and function as organizational memory. (Buckingham Shum, 1997; Pallotta et al., 2005)
- The development of argument diagrams may teach critical thinking. (Reed and Rowe, 2001; Van Gelder, 2003)

It is obvious that they can serve very similar functions when applied to records of meetings. Even more, if we can assess these argument diagrams automatically, they could be used in (pro-active) meeting assisting systems. The Argnoter system (see Section 1.4) would, for example, no longer require a manual operator.

6.2.1 Methods and Models

Several methods and models have been developed to structure argumentation in a way similar to the way we aim to realize for argumentation as it evolves in meeting discussions. All of these methods have their own goals in mind, their own ways of creation and hence their own benefits and drawbacks. We will discuss some of them here.

Wigmore's charting method Wigmore (1931) developed a graphical method for charting legal evidence, in order to make sense of a large body of evidence.

The purpose of his charting mechanism is to represent proof of facts in evidence presented on either side of a trial, to make sense of a large body of evidence. His charts depict the arguments that can be constructed from this body of evidence as well as possible sources of doubt with respect to these arguments.

In his model each arrow represents an inference or a provisional force. The nodes are the *facts* or the kinds of evidence that are put forward. Each type of evidence has its own shape. Circumstantial evidence is, for example, represented by a square, whereas testimonial evidence is represented by a circle. Furthermore there are possibilities for including a type of relation between facts where one fact 'explains away the other', whether the evidence was offered by the defendant, or whether the fact was observed by a tribunal or judicially admitted.

His diagram can reveal the logical structure of evidential reasoning in a powerful way and therefore it can be useful for automating legal argumentation of the kind prominent in law. Shum (1994) stressed that an important aspect of Wigmore's method is that his way of charting is not an attempt to express reasons of belief, but to express reasons of doubt. This way it reveals the weak points in the argument chain. Some regard Wigmore's charting method as a forerunner of theories of defeasible argumentation (Prakken et al., 2003). As the nodes can be interpreted as propositions, the vertical links as expressing defeasible inferences and the horizontal links as being relations of attack or defeat between arguments.

The Toulmin model In the late 1950's Stephen Toulmin developed a model that presents a 'schematic representation of the procedural form of argumentation' (Toulmin, 1958). That is to say, the role played by verbal elements in the argumentation during the justification process. Toulmin's model is concerned with pro argumentation and the acceptability of a claim.

Toulmin regards an argument as a sequence of interlinked claims or reasons that between them establish the content and force of the position for which someone is arguing. He states that an argument consists of six building blocks: a *datum* which is a fact or an observation, a *claim* related to the datum through a rule of inference which is called a *warrant*, a *qualifier* which expresses a degree of certainty of a claim, a *rebuttal* containing the allowed exceptions and a *backing*, which can be used to support a warrant.

Toulmin structures provide an intuitively plausible set of categories and relations for representing the logical structure of arguments organized in a distinctive graphical layout(Newman and Marshall, 1991), but on the other hand it has been suggested that his model does not lead to the production of unique representations for given arguments (Cooley, 1959) and that, as Toulmin's model is only concerned with pro argumentation, the model is inappropriate for depicting everyday arguments (Willard, 1976).

The IBIS model The IBIS model (Kunz and Rittel, 1970a) captures argumentation in terms of issues and their alternatives that have been proposed and

accepted by the participants (Note that IBIS is not a graphical diagramming model). It is based on the principle that the design process for a complex problem is a conversation between the participants who each have their own area of expertise.

IBIS is used to solve problems by using argumentative processes in a way to apply a structure to a problem. In the process the problem is also called the topic. Within this topic, speakers bring up issues. Whenever speakers have an opinion about an issue, they can assume a position to state how they look at the issue. To defend their opinion about the issue they can construct arguments until the issue is settled. In this process the participants give their opinion and judgement about the topic and thus create a more structured view of the topic and its possible solution (Conklin and Begeman, 1988). A serious drawback of this approach is that the labels that can be assigned to the nodes (text segments) are determined by a meta schema that predefines which relations are allowed to be attached to a particular node label.

Rhetorical Structure Theory Rhetorical Structural Theory (Mann and Thompson, 1987) is a theory of text organization. Rhetorical Structure Theory, or RST, was created with the intention to guide computational text generation (Taboada and Mann, 2006). RST addresses text organization by means of relations that can hold between the sentences in a text. It explains coherence by postulating a hierarchical, connected structure, in which every part of a text has a function or role, in relation to the preceding or the following part of the text. Some of the relations proposed in RST are: evidence, background, elaboration, contrast, condition, motivation, concession, restatement. The resulting RST trees are in theory capable of visualizing the rhetorical structure of a text, so with the right labels for the relations, the argument structure might be visualized. RST however lacks the ability to assign labels to fragments of a text.

All of the methods described serve their own purpose and show differences in application domain and in level of detail. What they have in common is that they all have labels that structure parts of discourse in such a way as to facilitate comprehension and to point out possible flaws. Our annotation schema should be able to reveal similar structures, not from written text, but from meeting transcripts. This goes along with the fact that not all arguments will be in favor of a particular issue and neither is it to be expected that all the components, such as defined by the Toulmin model, will be present, nor that all arguments will work towards a final conclusion.

6.2.2 Diagramming tools

Nowadays several computer software tools are available that are able to assist with the creation of an argument diagram. These Computer Supported Argument Visualization (CSAV) tools, or applications, are designed to assist in sorting and making sense of information and narratives found in minutes or other forms of discourse. Users of the tools are able to manipulate, annotate

and display the structure in various ways. All these tools provide means for the creation of an argument diagram, and all of them have their own underlying model or method with their own set of components from which, in the end, the resulting diagrams can be created. The components, or objects and relations, and the rules for combining them have been referred to by Suthers (2001) as 'representational notation', but in fact all these notations can be considered as annotation schemas and are therefore also examined for useful properties.

Most of the tools aim to provide a means for students and scholars in argumentation to analyze the structure of natural argument. Araucaria (Reed and Rowe, 2001), named after a tree, is for example such a tool. In Araucaria argument premises are to be placed below the conclusions and all nodes (propositions) and the connections between them can be labelled according to their evaluation. Another educational tool, that aims to increase critical thinking, is Reason!able (van Gelder, 2002). The primary objects in Reason!able are claims, reasons and objections. These components can be used to model argument trees. In the resulting argument trees, a 'child' is always evidence for or against a parent. Similar trees can be constructed with software packages such as Athena[2] and Belvedere (Suthers et al., 1995).

A somewhat different tool is Compendium (Selvin et al., 2001), which was designed as a tool to support the real time mapping of discussions in meetings, collaborative modelling, and the longer term management of this information as organizational memory. Another difference with the other tools is that the resulting diagram can contain, apart from arguments or conclusions, questions or issues as well as answers or ideas that have been expressed. Furthermore decisions can explicitly be indicated and references to external data sources can be included such as notes and spreadsheets.

There are some differences between the capabilities of these tools. Araucaria is for instance able to handle argumentation schemes in such a way that if a complex of propositions is a joined structure, the whole structure can be labelled. In Athena, users are able to manually assign a relevance value to the relations and to manually evaluate the acceptability of the premises to see how much strength a parent would derive from its children. In Reason!able one is able to evaluate arguments on their strength (on a three level scale: no support, weak support and strong support), the degree of confidence in their truth, and on independent grounds for accepting or rejecting (e.g. because it was stated by an authority). The Belvedere environment allows the nodes to be labelled with labels as *Principle, Theory, Hypothesis, Claim, Data* where as in Reason!able, the nodes can be only of type *Claim*. It appeared that the positive (support) and negative (refute) relation between arguments are included in all of the examined tools. Only in the Belvedere environment are the relations somewhat finer grained: examples of their relation set are *support, explain, undercut, justify, conflict*. Another observation is that in all of the tools, except compendium, the main conclusion or thesis that was debated is represented as the uppermost node.

[2] www.athenasoft.org

6.3 Developing the Annotation Schema

An argument structure is best seen as a reconstruction of a sequence of reasoning. In order to automatically obtain an argument structure, a model, or annotation schema, is required that initially can be manually applied by humans. The creation of such a schema is, in contrast to the model in the previous chapter,a much less straightforward enterprize. One of the findings, with respect to all diagramming models that were studied, was that they generally start with, or work towards, a final 'conclusion'. In the domain of meeting discussions where people make decisions, however, there might be no conclusion at all (e.g. due to time constraints). The main aim therefore is to capture contributions, or parts of contributions, regardless of whether consensus is reached. The resulting structure is then to provide insights into the issues debated and the statements made.

We intend to use a graphical representation of argumentation, that is comparable to the argument diagrams that were presented in the previous section. This way, the resulting structure itself could instantly provide some notion of how the discussion took place. The schema was to capture information closely related to the kind of relations found in the model diagrams described in the previous paragraph, but also required the participants' utterances, or units, to be labelled. This unit labelling relates highly to Dialogue Act labelling. Dialogue acts are labels for utterances which roughly categorize the speaker's intention. (see Bunt (1979)) and relate to speech acts as described by Austin (1962) and Searle (1969). DA labels serve as elementary units to recognize higher levels of structure in a discourse. An example of a dialogue act scheme is for example described in Jurafsky and Shriberg (1997) (See also section 6.5).

So, our schema should contain labels for the utterances indicating their argumentative function in the discourse and contain labels that indicate the argumentative relation that holds between them. But what labels should it have, and what constraints are to be taken care of? The number of labels that an annotation schema has, sometimes referred to as the 'depth of the palette', is generally a trade-off between expressivity and ease of use (Selvin, 2003). According to Bruggen (2003) the most important question that needs to be answered, before one starts to define an annotation schema is *what* must it contain? In our case the schema should visualize the structure of our design meeting discussions containing the contributions from the meeting transcripts in a crisp and coherent way, such that answers to questions asked about the discussion either follow directly from the discussion schemas or can be derived in a straightforward and easy manner.

Walton and Reed (2003) describe five what they call 'desiderata' for models describing the components and the relations between these components in order to constitute an argumentation diagram. The desiderata are:

1. Rich and sufficiently exhaustive to cover a large proportion of naturally occurring argument.

2. Simple, so that they can be taught in the classroom, and applied by stu-

dents.

3. Fine-grained, so that they can be useful, and employed both as normative and evaluative system.
4. Rigorous, and fully specified, so that they might be represented in a computational language.
5. Clear, so that they can be integrated with the traditional diagramming techniques of logic textbooks.

These desiderata also seem useful to apply to our annotation schema.

Our to-be developed schema is, however, faced with one drawback that is inherent to argument diagramming and argumentation in general, that is, that there is no correct diagram. Walton (1996) for instance showed that various different argument diagrams can be instantiated by one single text. This implicates that there can be more than one reasonable analysis, which makes it hard, if not impossible to evaluate the annotations that are created with the schema in terms of right and wrong. Although this is a problem, Reed and Rowe (2001) point out that an analyst in this case should always make *plausibility judgements* rather than absolute analytical decisions. Or to quote from Polyani (1988) "As long as our interlocutors believe that we have assigned a plausible interpretation to their remarks, they are not disappointed if that interpretation does not exactly reflect what they had in mind." This in essence shows that the observer is free to interpret and to create that diagram that he, or she, considers the most appropriate according to his or her own perception. As long as the schema is applied correctly, its purpose will be apparent anyhow. It may even be a good idea to have alternative diagrams to represent two or more possible interpretations. As long as the argument diagram is useful to present a visual summary of the flow and direction of the argument.

6.3.1 The Twente Argument Schema

The Twente Argument Schema (TAS) is an annotation schema designed to define argument diagrams for meeting discussion transcripts. Following most of the existing diagramming techniques, application of the method results in a structure with labelled nodes and edges. The nodes of the tree contain complete speaker turns or parts of speaker turns whereas the edges represent the type of relation between the nodes. The complete label set is shown in Table 6.1.

It has been tried to identify the various functions of the argumentative aspects of the different contributions made by the participants and also to define labels that relate these contributions to each other. The approach that was taken was a so-called 'goal driven design' approach. Based on the literature on argumentation theories and argument diagramming, argument diagrams were created on a small corpus. In several rounds we tried to reach a consensus on how to chart a meeting discussion, in terms of structure and the associated labels. When required, the representational notation was refined. The whole process was repeated until agreement was reached on the labels for the components.

Node labels	**Relation labels**
Statement	Positive
Weak statement	Negative
Open issue	Uncertain
A/B issue	Request
Yes/No issue	Specialization
	Elaboration
	Option
	Option exclusion
	Subject-to

Table 6.1: The labels of the Twente Argument Schema

It was quickly decided to start to work from manually created transcriptions. This choice was made because the state-of-the-art in speech recognition does not yet result in good meeting transcripts. The automatic creation of meeting transcripts is an isolated problem, that one day could be solved and is therefore left behind in our efforts to automatically create an argument structure from the meeting transcripts. As meetings unfold in time, it was furthermore decided to construct our model in a sequential order that follows the line of the discussion. This way the layout also facilitated comprehension (cf. Bateman et al. (2001)).

It was also decided to work with graphs in the form of trees[3]. One of the main reasons to do so, was that TAS this way could preserve the conversational flow. (By applying a left-to-right, depth-first walk through the resulting trees, the reader is able to read the nodes as they unfolded in time.) This was realized by assuring that in principle every next contribution of a participant becomes a child of the previous contribution, unless the current contribution relates more to an ancestor. Another advantage of the fact that TAS trees are created in a left to right manner, is that they end up with a relatively small set of attachment points, or fringe (cf. the Discourse Parse Trees created by Polyani (1988)).

An example of a TAS argument diagram, embedded in a meeting browser application, is shown in Figure 6.1.

[3]See Baldridge and Lascarides (2005) for an elaborate discussion on this topic.

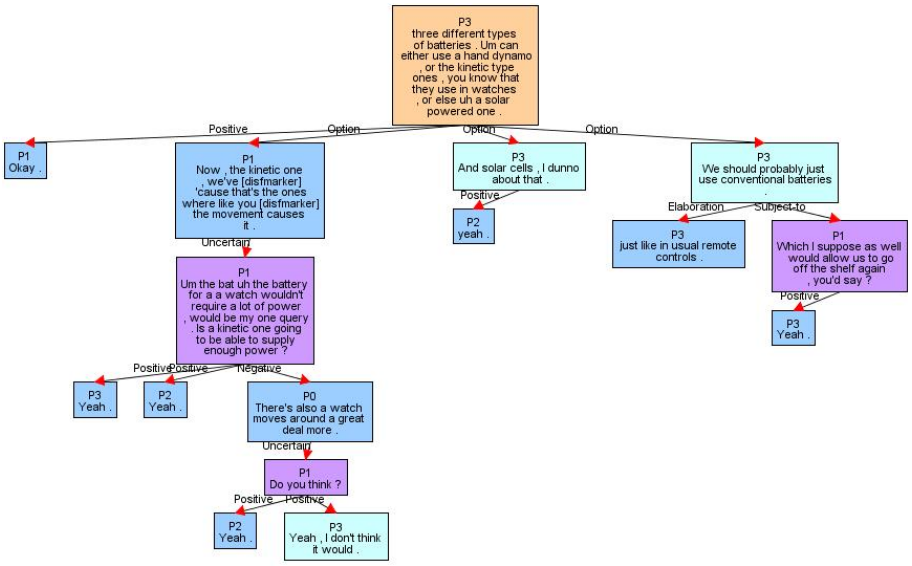


Figure 6.1: An example of a TAS Argument Diagram.

6.3.2 The Unit Labels

The content of the nodes correspond in granularity to the size of speech acts, resulting in most instances in the size of a complete utterance. If utterances contain more than one act, they are split up into more than one node. In line with Galley et al. (2004) backchannel utterances such as 'uhhuh' and 'mmhm' are filtered out, since they are generally used by listeners to indicate they are following along, and not necessarily indicating (dis)agreement. The nodes in TAS consist of issues and statements.

Issues can also be found in the IBIS model (see Kunz and Rittel (1970a)). There, they are represented as questions as they can be seen as utterances with a direct request for a response. Kestler (1982) distinguishes two fundamental types of question with respect to conversational moves. These are *yes-no questions* and *why questions*. A yes-no question admits only two kinds of answer, be it either supportive, or negative but rules out the uncertainty *option* 'I don't know'. The *why questions* are a subclass of a more general type of *open question*. The number of positions participants can take on such an issue depends on the set of possible options enabled by the type of question or issue.

In our scheme we have defined three different labels for our nodes to represent the issues: The **'Open issue'**, the **'A/B issue'** and the **'Yes-No issue'**. The open issue allows any number of possible replies possibly revealing positions or options that were not considered beforehand. This in contrast with the A/B issue, that allows participants to take a position for a number of positions which should be known from the context (c.f. 'Would you say ants, cats or cows?').

The yes-no issue, in line with the yes-no question directly requests whether the participants' positions are in line with, or contradict the issue. A *why question* in TAS is modelled as an open question with a clarification relation (see below).

The positions that participants take are generally conveyed through the assertion of a **Statement**. The content of a statement always contains a proposition which can be a description of facts or events, a prediction, a judgement, or an advice (Van Eemeren et al. (2002)). Statements can vary in their degree of force and scope. Meeting participants may indicate that they are not sure if what they say is actually true. In Toulmin (1958) *qualifiers* provide an indication of the force of *claims*. As Van Eemeren (2003) points out, the force of an argument can also be derived from lexical cues such as the words 'likely' and 'probably'. Such statements, in which the speaker does not commit himself fully to the opinion are labelled as **'Weak Statements'** in TAS. All the remaining transcription segments that could not be labelled with any of the above labels were labelled as **'Other'**.

6.3.3 The Relation Labels

From several intellectual subfields various researchers have produced lists of inter-segment relations from philosophers: (e.g. Toulmin (1958)) to linguists (e.g. Quirk and Greenbaum (1973); Halliday and Hassan (1985); Martin (1992)) to computational linguists (e.g. Hobbs (1979); Mann and Thompson (1988b)) to Artificial Intelligence researchers (e.g.Schank and Abelson (1977); Dahlgren (1988)). The general approach contains somewhere between five and fifty different relation types. For TAS, we have defined nine relations that can exist between the labelled nodes. These are described below.

When engaged in a discussion or debate, the elimination of misunderstandings is a prerequisite in order to understand each other and hence to proceed (Neass, 1966). Participants in a discussion, according to Neass, eliminate misunderstandings by generalizing, or specifying their statements. The **'Specialization'** label was therefore introduced. This label can be applied when a particular issue generalizes or specializes another issue. The contribution 'Which animal is the most intelligent?' can be specialized with the following contribution 'Is an ant or a cow the most intelligent animal?' which again can be specialized if one for instance asks 'Are ants the most intelligent animal?'. It is also possible that a person is not satisfied with the information or the argument explained. He can then explicitly invite the previous speaker to elaborate on his earlier statements. For these situations we define the relations **'Request'**. The **'Elaboration'** label is used if a person continues his previous line of thought and adds more information to it.

Whenever an issue is raised, an exchange of ideas about the possible solutions occurs in the decision making process. As questions call for answers, issues call for opinions expressed through statements. Whenever a statement is made as a response to an open-issue or an A/B-issue it might reveal something about the opinion of the participant on the solution space. In general a participant provides an **'Option'** to settle the issue at hand. For example when a speaker

asks 'Which animal is the most intelligent?' and the response from someone else is 'I think it's an ant' the option relation is to be applied. The opposite of the option relation is the **'Option-exclusion'** relation, and it is to be used whenever a contribution excludes a single option from the solution space.

With respect to a yes/no-issue the contributions that can be made are not intended to enlarge or to reduce the solution space, but to reveal one's opinion to the particular solution or option at hand. Contributions from participants are either supporting, or objecting to the issue, or express uncertainty. For this purpose the labels **'Positive'**, **'Negative'** and **'Uncertain'** are introduced. The positive relation can exist for example between a yes/no-issue and a statement that is a positive response to the issue or between two statements agreeing with each other. When one speaker states that cows can be eliminated as being the most intelligent animals and the response from another participant is that cows don't look very intelligent, then the relation between these statements is positive. The negative relation is to be applied in situations where speakers disagree with each other or when they provide a conflicting statement as a response to a previous statement or a negative response to a Yes/No-issue. In a case where it is not clear whether a contribution is positive or negative, but that there exists some doubt on the truth value of what the first speaker said, the uncertain relation is used.

The final relation of our set is applied when the content of a particular contribution is required in order to figure out whether another contribution can be true or not. We termed this the **Subject to** relation. It is related to the concession relation in Toulmin's model. It is applied for example in the situation where someone states 'If you leave something in the kitchen, you're less likely to find a cow' and the response is 'That depends on whether the cow is very hungry'.

6.3.4 Corpus Creation and Reliability

An annotation tool called *ArgumentA* was created for the annotation task by building further on a number of components described in Reidsma et al. (2005a). ArgumentA allows annotators to select text on a transcription-view panel and label them, similarly to dialogue-acts. The label is assigned by selecting the unit text with the mouse from the transcription panel and then pressing a button popping up a label selection window from which the unit label can be picked. The labelled units appear on a canvas where they can be attached to the graph via an intuitive drag and drop interface. Once attached, a popup window appears from which the relation-label can be chosen. The resulting trees could be saved in both NXT-format and in a specific XML format designed for this scheme. The latter we used as input for our classifiers described in Sections 6.5 and 6.6, whereas the former is used in, for example, the browser plug-in.

Three annotators were trained in several iterations. Apart from collectively developing the scheme, elaborate discussions were held after a number of training sessions about when and why to pick a particular label in that particular case. To measure the reliability of the scheme we compared the unit labels on

pre-segmented discussions for four meetings (12 discussions) between two of our annotators. It turned out that, especially in first trials the value of Cohen's kappa (κ) Cohen (1960) were rather low (0.50) as there was a lot of confusion about the labels 'other' and 'statement'. This was resolved by a consensus definition for the word 'yeah', after which κ rose to a more acceptable value (0.87). In contrast, but not comparable, to our results Nomoto and Matsumoto (1999) report a κ of 0.43 on RST annotations on texts using three naive coders. Carlson et al. (2001), on the other hand, report a κ in the range between 0.95 and 1.00 on unit segmentation and a κ in the range between 0.62 and 0.81 for labelling predefined relations using professional language analysts who went through a long period of extensive training. In earlier work Marcu et al. (1999) reports κ's between 0.73 an 0.79 on segmentation and 0.53 and 0.64 on relation labelling. We did not compute *kappa* scores on our relation labelling task, because annotators identified and labelled the relations in one single run.

With respect to the issue of reliability one should note, as mentioned, that for this task it is very easily possible to end up with several diagrams from one discussion as there are likely to be more than one possible interpretation. Walton (1996) for instance showed that various different argument diagrams can be instantiated by one single text. Moreover, in Rhetorical Structure Theory (RST) (Mann and Thompson, 1988a), which addresses similar issues as the TAS scheme, the suggestion is made that the analyst should make *plausibility judgements* rather than absolute analytical decisions, implying that more than one reasonable analysis may exist. Carletta (1996) in this respect even states that in subjective codings such as these in the case of argumentation, there exist no real experts and that the only thing that counts is how totally naive coders manage based on written instructions.

After the trials were finished and all the annotators were convinced of the fact that they were able to apply the scheme in the way that it was intended to be used, an annotation manual was created and all the meetings from the AMI Corpus (See Section 1.3.1) were manually annotated over a period of three weeks accordingly. Each of the three trained annotators annotated one third of the whole corpus. The final distribution of TAS unit and relation labels on the AMI corpus is shown in Table 6.2.

As we now had an annotated corpus that was to be used as training data for classification experiments, the fact that the reliability measures were computed on just four meetings was a bit dissatisfactory. To overcome this, an alternative view on the reliability scores, that could be called the *virtual* κ, was calculated. This is a technique that is comparable to what is introduced in Steidl et al. (2005) and requires the feature values that one will use for the classification task. These features are explained in Section 6.5 and Section 6.6.

The idea behind virtual κ is that one sets out the results of a classifier trained on annotations of one observer against the class labels provided by another annotator. The maximum and minimum performances this way do not just provide an indication of the interval in which the average performance of the classifier might lie, but the difference between the minimum and maximum performance also gives an indication of the similarity of the individual annota-

Node labels	Amount	**Relation labels**	Amount
Statement	4077	Positive	2319
Weak statement	194	Negative	471
Open issue	232	Uncertain	259
A/B issue	69	Request	223
Yes/No issue	443	Specialization	131
Other	1905	Elaboration	689
		Option	601
		Option exclusion	14
		Subject-to	190
Total	6920		4897

Table 6.2: Distribution of TAS labels

tions. The more they look alike, the more the classification performance of the values that result from training and testing on the same annotator resemble the values that result from classifiers that were trained and tested on a different annotator.

The performances shown in Table 6.3 show the results for training and testing the classifier on the unit labels. When both training and test sets were picked from the same annotator, we used 10-fold cross-validation.

Trained / Tested on	Annotator 1	Annotator 2	Annotator 3
Annotator 1	84.4%	75.7%	70.3%
Annotator 2	75.6%	79.5%	66.2%
Annotator 3	67.0%	66.2%	82.2%

Table 6.3: Performance on unit labels amongst annotators

For Table 6.3 it thus appears that the annotation from Annotator 3 differs more than the annotations provided by Annotator 1 and 2. This was even more clear on the table we obtained for the relations. See Table 6.4.

Trained / Tested on	Annotator 1	Annotator 2	Annotator 3
Annotator 1	59.80	58.10	44.70
Annotator 2	62.53	64.77	53.31
Annotator 3	40.20	41.47	50.18

Table 6.4: Performance on relation labels

In order to increase the classification performance, one can, based on these tables, decide to use just those annotations that look alike. However, a trade-off exists here again. Because, although the performance could increase this way, a model that is trained on more annotators, is also likely to better represent the general interpretation of the annotation scheme.

6.4 Related work on automatic argument diagram creation

Several steps can be distinguished on the road towards full automatic argument diagram creation. Given an annotation schema, such as TAS, that contains the labels for the observations that are to be recognized, the steps towards full construction from the moment that transcriptions become available to the system are the following:
1) The meeting is to be segmented into discussion/non-discussion segments.
2) The discussions are to be segmented into units that can be labelled.
3) The detected units need to be labelled appropriately.
4) The relations between the labelled units have to be determined, and
5) The detected relations need to be labelled appropriately.

The research described in the remaining sections is confined to the classification tasks of steps 3 and 5 and assumes pre-selected discussions, unit segments and relations between unit segments to be available. This section intends to give an overview of the literature and techniques that are available and that approach the issues that we considered given for our practice.

6.4.1 Automatic Discussion detection

Discussion detection in meetings has been described as a subproblem of meeting activity detection. Meeting activity detection deals with segmentation of meetings into a number of group activities that usually correspond to location-based turn-taking patterns, including monologues, discussions and presentations (Gatica-Perez, 2006). Detected sequences of meeting activities can be used to provide a summary of the meeting structure. Automatic detection of human interactions generally uses low-level and usually multi-modal signals as input. Although approaches exist that focus on single modality features such as audio (Dielmann and Renals, 2004), better results have been obtained with a multi-modal approach that also considers video, as well as audio features (McCowan et al., 2005; Al-Hames and Rigoll, 2005; Al-Hames et al., 2005; Reiter et al., 2006; Zhang et al., 2006).

Audio features comprise, for example, speech/silence segmentation, prosodic information and speaker activity, whereas video based features typically contain head and hands location, eccentricity and motion direction (Al-Hames et al., 2005). Feature values have generally been calculated for each of the participants individually. Recent approaches however incorporate the notion from McGrath (1984) that meetings contain both individual actions as well as group level interactions. As a consequence also group level features such as usage of the whiteboard or data projector have been used (Zhang et al., 2006; Al-Hames et al., 2005). This trend is also observable in Zhang's model that was used for dominance hierarchy detection, as described in Section 5.3.

Computational models that predict sequences of meeting actions that take both individual and group level features into account are so called two-layered

Hidden Markov Models (HMM's) [4] . The first layer maps the low level features to individual actions and the second layer uses results from the first layer to recognize group actions. The approach followed by Zhang et al. (2006), for instance, yielded a performance of about 70%, whilst using a set of 15 distinct meeting activities and a 59-meeting corpus.

Our needs however differ somewhat from the works described above in a way that for our purpose a binary classification of meeting actions into discussion and non-discussion would suffice. The question as to whether this segmentation can be carried out with a higher precision remains to be investigated. The task may seem easier with just two class labels on the one hand, but on the other hand we read in Zhang et al. (2006) that they especially noticed difficulties in discerning monologues from discussions.

Other literature with respect to meeting segmentation can be found in the area of topic segmentation (Hsueh et al., 2006; Galley et al., 2003), text-tiling (Hearst, 1997), and 'meeting hot-spot detection' (Wrede and Shriberg, 2003b,a). Hot spots are meeting segments with highly involved participants that are claimed to be relevant for browsing and retrieval purposes in contrast to segments containing low involvement. All of the approaches described could prove beneficial for our aims.

6.4.2 Automatic Unit segmentation

The automatic identification of units in essence boils down to the segmentation of the discussions mentioned above into segments or chunks amongst which relations exist. Research on discourse segmentation has relied on various definitions of discourse segments. Discourse segments have amongst others been defined in terms of participant's intentions (Grosz and Sidner, 1986), in terms of an informal notion of topic (Hearst, 1997) and in terms of subdialogues that accomplish one major step in the participants plan for achieving a task (Carletta et al., 1997).

Passonneau and Litman (1997) have shown that humans agreed on segment boundary identification when they apply the intention based definition of Grosz and Sidner (1986) on a corpus of spontaneous, narrative monologues. The best algorithm that was to replicate these intention based discourse segments recalled 53% of the discourse segments identified by humans with a precision of 95%. The features that were used contained manually encoded linguistic and non-linguistic features that pertained to prosody, cue phrases and referential links.

Marcu (1997a,b) proposed a surface based approach that relies primarily on cue phrases and so-called lexicogrammatical constructs that can all be detected without a deep syntactic and semantic analysis. An annotated corpus was used to semi-automatically deduct rules for segmentation. When markers with their orthographic environment (commas, periods, dashes, etc.) in more than 90% of the corpus occurrences identified a new segment (e.g. "{,} although"), a rule was instantiated marking the boundary of the segments it connects. For

[4] For more information about (types of) Hidden Markov Models, see Murphy (2002)

the markers with an associated rule a second rule was created that informed the system on how to determine the borders of the textual units to which the marker, or cue phrase, belongs. For "{,} although" the associated rule states that the textual unit starts at the marker and ends at the end of the sentence, or at a position determined by the rule from the next cue phrase in that sentence. The resulting algorithm found 80.8% of the discourse markers in a test set. This in comparison to, for example, Hirschberg and Litman (1994) who have experimented with prosody and pith (75.4%) and POS information (63.9%). Marcu's method recalled 81.3% of the boundaries that were agreed upon by at least two of three judges who were asked to intuitively break-up three different texts into units.

It should be said that orthographic information was provided for this task. In our case, where we also work with transcripts, this information is present as well, but when one whishes to find segments fully automatically from speech signal input this annotation layer should be eventually automatically deducted from the speech signal itself. For advances in this direction consider the work of Kim (2001) for a prosodic approach, and the work of Huang and Zweig (2002) for a MaxEnt approach.

It remains, however, to be investigated how well these methods transfer to transcripts, and in our case to unrestricted multi-party speech. A clear distinction between the multi-party speech and text is that the transcriptions resulting from automatic speech recognizers require additional post-processing in order to become to some extent 'comparable'. This requirement is due to the fact that during multi-party speech issues such as disfluencies, coarticulation, word fragments, and ungrammatical utterances arise. All of these are in turn problematic for current state-of-the-art speech recognizers (Liu, 2004). This post processing, sometimes also referred to as summarization (See e.g. Zechner (2002)), has proven not just to increase parsing results (Kahn et al., 2004), but also to increase readability (Jones et al., 2003); both of which are important for argument diagram creation. On the other hand, one could expect to end up with better performance for automatic segmentation of multi-party speech than for text and single speaker speech (see also (Hirschberg and Litman, 1994; Marcu and Echihabi, 2002)). One reason for this could be the fact that speaker changes explicitly mark segment boundaries. An initial experiment with the TAS corpus indeed shows that the start-times of 77.04% of the TAS unit labels coincide with a start-of-speech border taken from the speech/silence segmentation that was created to aid the AMI transcription process (Lathoud et al., 2004; Moore et al., 2005). On the other hand, as mentioned above, orthographic information that has proven to aid the segmentation process is initially absent in ASR output.

6.4.3 Automatic Relation detection

Any reader can distinguish a text or a transcript from a random set of sentences. This is due to the fact that human readers are able to relate (part of) sentences as belonging together in some way and in some form. Research on discourse structure has attempted to determine and typify these relations. It has become

widely accepted that sentences are connected to each other by means of two linguistic phenomena, namely cohesion on the micro-level and coherence on the macro-level. Coherence relations refer to a relation between sentences that contributes to their sense making (Morris and Hirst, 1991). One speaks, on the other hand, of cohesion when the full comprehension of an element in a given discourse is dependent on another preceding element. Cohesion in this sense can therefore be regarded as an objective discourse property, whereas coherence is produced by the evaluation of the readers trying to capture the writer's or speaker's intention.

Although coherence relations come closest to the relations in the TAS schema, cohesion relations can be used to identify coherent parts of the discourse (See e.g. Barzilay (1997); Cristea et al. (1999)). We therefore first chart the most important cohesion relations together with pointers to literature that apply techniques for their detection, before zooming in on the issue of automatic coherence relation detection.

Cohesion relations

Cohesion relations can be divided into four categories: reference, conjunction, ellipsis and lexical cohesion relations (Halliday and Hassan, 1985).

When an item, such as a word, is linked by a semantic relation to some element in the preceding text we speak of a *reference*. For an example of cohesion through reference consider these two sentences: *"Thomas was given a bike. He liked it very much."*. The anaphoric expressions *He* and *it* refer respectively to the antecedents *Thomas* and *bike*. As a consequence they relate the two sentences. The problem of (automatically) determining the proper antecedent of a given anaphoric expression in the current or the preceding utterance(s) is known as anaphora resolution. A good starting point for work in the area of anaphora resolution is Mitkov (2002). Schauer (2000) exploits cohesive reference relations to derive *referential constraints*. Resolution of anaphoric expressions is used to limit the set of potential target nodes. This approach is said to correctly predict 86.4% of the target units.

Cohesion through *Conjunction* expresses logic-semantic relations between clauses explicitly. The word 'because', for instance, explicitly signals a causal relation. According to Schauer (2000) between 15-20% of all coherence relations is explicitly signalled by a connective such as 'and' or 'or'. Conjunction words are often explicitly embedded in Part-of-Speech categories and lend themselves for easy and straightforward detection.

Cohesion through *ellipsis* occurs when a clause can be presupposed, but is omitted. Consider: *"Guido was walking to the school and then to his house"*. The verb 'walking to' explicitly relates to 'school' and implicitly to 'house'. For computational approaches to detect ellipsis, see, for example, Carberry (1989) and Schiehlen (2002).

Lexical cohesion, finally, occurs through the usage of words that relate in some way to words that have been used before. Halliday and Hassan (1976) distinguish two classes of lexical cohesion: Reiteration and Collocation. Reit-

eration occurs when a lexical item brings to mind the meaning of a previous item. This can be achieved through repetition, the usage of homonyms and synonyms, the use of part/whole and whole/part relations and through word association through specialization or generalization. Collocation refers to words that co-occur in discourse. The words 'cow' and 'farm', for instance, can generally be expected to co-occur more frequent than the words 'cow' and 'kitchen'. Lexicons such as WordNet (Miller, 1995) are typically used to calculate distances between words. If lexical cohesion occurs between a sequence of words one speaks of a *lexical chain*. An approach to the automatic detection of lexical chains is described in Barzilay (1997).

Coherence Relations

Coherence relations, also called discourse relations, or rhetorical relations, are said to realize the intentions of the speakers or writers, with the idea that readers or listeners will recognize them when interpreting the discourse (Grosz and Sidner, 1986; Taboada and Mann, 2006). There are theories that state that only a single relation may hold between syntactic clauses, or segments (Mann and Thompson, 1986). Others have stressed the need for more than one layer of relations Moore and Pollack (1992), or even several layers with crossing dependencies (Webber et al., 1999). We confine ourselves to single-layered approaches due to computational and visualization issues.

A pragmatic approach, when wishing to identify relations between units is that one for each (source) unit starts with hypothesizing a number of (target) units to which the source node potentially is related. Without any knowledge about the type of node, one can make use of the fact that when discourse units are placed adjacent to one another, that is, come after another, people are likely to infer a relation between the two (Webber et al., 1999). To investigate how well this heuristic performs on the TAS corpus we consider the distance between all the related units in the TAS corpus (See Figure 6.2).

It appears that most of the relations in the TAS corpus (47.54%) are relations between subsequent fragments of discourse.

According to Corston-Oliver (1998a) three strands have emerged in the field of computational approaches to coherence relation detection. The first strand concerns the form of the text. Usually identification proceeds by fairly superficial means such as pattern matching with regular expressions (see e.g. Marcu (1997a); Kurohashi and Nagao (1994)). The second strand, considers more abstract representations. Lexical items are here to be augmented with axiomatic representations of world knowledge, such as described in Hobbs (1979). It is assumed that structures can be built in conjunction with fully specified clauses and sentence typologies (see e.g. Lascarides et al. (1992)). The third strand concerns a more programmatic description of how computational discourse analysis might proceed. The broad strokes of the design of a computational discourse analyzer are described, but no specific details are given (see e.g. Polyani (1988)).

It appears that all, except the first strand goes beyond the current state of technology, as they are too informal and do not support a procedural approach

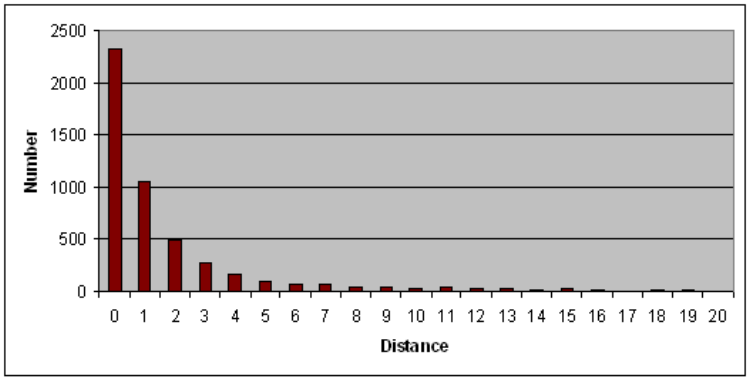


Figure 6.2: An overview of the number of units between related units in the TAS corpus.

that can be automated. Some examples of the first strand are discussed in Section 6.6.

6.5 Assigning TAS unit labels to predefined segments

In this section we report on experiments related to the automatic labelling of speech segments with TAS-unit-labels. First, in order to give an overview of related literature and features that have been used for a similar task, related work in the area of dialogue act classification is discussed, then the features are described that we used as input for our classifier, before we end this section with describing our obtained performances.

6.5.1 Related Work

The labelling of text segments with TAS unit labels, or any other set of (class) labels shows much resemblance to what is called dialogue act tagging. Dialogue acts (DA's) are, as mentioned, labels for utterances which roughly categorize the speaker's intention. The task to assign a DA out of a predefined set of possible dialogue acts to an utterance is called dialogue act tagging. And although our goal is to model the argumentative function of the utterances of the speaker, rather than to assign labels that exhaustively try to enumerate people's intentions, both classification tasks are in fact similar (cf. Verbree et al. (2006)).

The topic of automatic dialogue act tagging has received quite some attention in the past years (cf. Jurafsky et al. (1998); Rotaru (2002); Shriberg et al. (2004)). A variety of methods has been tested on various corpora using different sets of dialogue act labels. The ICSI Meeting Corpus (Janin et al., 2003) is one

of the larger corpora and includes 75 naturally occurring meetings containing roughly 72 hours of multi-talk speech data and associated human generated word-level transcripts. It was hand-annotated for dialog acts as described in Shriberg et al. (2004) using the Meeting Recorder Dialog Act tagset (MRDA). The MRDA scheme has 11 general tags and 39 specific tags. Each annotation requires one general tag and a variable number of specific tags. Another renowned corpus is the Switchboard Corpus (Godfrey et al., 1992). This corpus consists of conversational speech by telephone. For a subset of this corpus, consisting of over 210,000 utterances grouped in 1,155 conversations, the dialogue act annotations based on the SWBD-DAMSL tagset are available (see Core and Allen (1997)). The best performance for DA classification on the ICSI corpus (81.18% accuracy) has been reported in Ang et al. (2005). For the Switchboard Corpus Rotaru (2002) reports an accuracy of 72%.

When we look at the features, three groups can be distinguished. In the first place there are language models that are based on words or part-of-speech tags. Second there are prosodic features like pitch and pitch slopes of the frequency spectra of the words, and also the duration of words, the vowels and the pauses. The third category of features are contextual features that describe the relation between the current and the surrounding utterances. Verbree (2006) gives a more elaborate overview of features that have been used in recent DA classification literature.

6.5.2 Features

A general assumption between scholars in the field of dialogue act classification is that the words and phrases in DA's are the strongest cues to their identity (cf. (Jurafsky et al., 1998)). Apart from some basic features most attention has therefore been given to create a useful, but small feature set created from N-grams of words and Part-Of-Speech (POS) tags. The full list of features that were used for the automatic assignment of TAS unit labels to classify utterances is shown in Table 6.5.

The fact whether a question mark token is present or not. (QMT)
The fact whether the word 'or' is present or not. (ORT)
The length (number of words) of each segment. (L)
The label of the previous two labels -manually assigned- (LL)
Is the current segment uttered by the same speaker as the previous one. (NS)
A vector of feature values for POS N-grams (P)
A vector of feature values for word N-grams (W)

Table 6.5: Features that were collected for the classification of TAS unit labels.

The approach that was taken to create the feature vectors for P and W was as follows: Each N-gram found in a test-instance was for each class compared to

a set of very predictive cue-ing N-grams. For each list on which the examined N-gram is listed the associated class was awarded a number of points. So the more N-grams of an instance (utterance) match a particular class, the higher the score for that class will be. If an instance, for example, starts with the Bi-gram 'I would' and appears to be listed in the top X for the classes 'A/B Issue' and 'Other', points are awarded to those two classes. Our resulting feature vector thus contains values for all the different N-gram orders (uni-, bi- and tri-grams) for each of the classes. The P and W features were devised to overcome the problem of huge feature vectors that contain binary values for each N-gram and dramatically reduce the computational speed.

Perl scripts were used to extract all of the features from the transcripts. The construction of N-grams was done using the N-gram Statistic Package (NSP) (Banerjee and Pedersen, 2003). The POS tagging was done with the Stanford Part-of-Speech tagger (Toutanova et al., 2003).

Tuning Parameters

An experiment was devised to derive the best vector of features for P and W. Four variables were put to a test in order to answer four questions: 1)How many N-grams should constitute the Top X of most predictive N-grams for a particular class? 2) Which ranker is to be used so that the N-grams can be ranked? 3)What number of points is to be awarded to the class whose N-gram is present in the test instance, and 4) Does pre-processing of the resulting feature vector scores have any positive influence on the performance?. A 2x2x2x2 experiment was created to get some insight into the possible answers. The number of N-grams that was used to constitute the top X predictive features for each class was set to be either 300 or 10. Two algorithms were used to rank the N-grams with respect to their cueing power to end up for each individual class with a top-N of uni-, bi- and tri-grams.

$$Pnts_{class} = \frac{\#Ngram^2_{class}}{\#Ngram_{allclasses}} \tag{6.1}$$

and

$$Pnts_{class} = P(\neg Ngram|class) + \sum P(Ngram|AllOtherClasses) \tag{6.2}$$

The first ranker (Equation 6.1) is mentioned in Verbree (2006) and the second (Equation 6.2) is the best performing ranker method mentioned in Samuel et al. (1999). The lists that resulted from the equations had to be sorted descending and ascending, respectively. The number of points, that was awarded each time an N-gram of the test-instance was found in a top X list for a specific class, was varied by either adding the number of occurrences of the N-gram in the class (in the training set) to the class total, or by adding just one single point to the class total. The final variable that was introduced is an additional pre-processing step in which, once all the points had been awarded to all the classes,

either the points were kept, or the value for the class with the most points was set to 1 and all the values for the other classes to 0 (per N gram type). This step was motivated by the fact that the decision tree learner that was used (J48) is unable to learn relations between features and this way the most predictive class is made explicit. Four orders of N-grams were used (Uni-, Bi-, Tri-, and Quadri-grams) for both word and POS N-grams.

As the collection of nodes in the TAS-corpus is rather small (<7k nodes), it was decided to conduct the experiment on the larger corpus of AMI Dialogue acts (>102k DA's). As stated, for both classification tasks an utterance is to be assigned a class label and the tasks are therefore more or less similar. The experiments were conducted making use of four distinct feature sets. The first two, I and II, made use of the L, P and W feature, whereas the other two, III and IV, made use of all the features. Another distinction between the feature sets, separating I and III from II and IV, was the way the top X feature list for each class was calculated. This was done either by making use of each order N-gram (OS) individually, or by summing the uni-, bi-, tri-, and quadri- grams (NOS). The results are shown in Table 6.6 and 6.7.

	Top 10							
	Ranker = Eq. 6.1				Ranker = Eq. 6.2			
	Pt = #		Pt = 1.0		Pt = #		Pt = 1.0	
	PP	¬PP	PP	¬PP	PP	¬PP	PP	¬PP
I	55.17	53.94	47.56	53.19	51.58	51.41	51.45	51.48
II	50.65	50.69	50.65	50.69	50.58	50.76	50.65	50.69
III	56.33	55.54	49.66	54.81	53.50	53.36	53.56	53.16
IV	53.21	53.19	53.21	53.19	53.16	53.01	53.21	53.18

Table 6.6: I = L_P_W (OS), II = L_P_W (NOS), III = QMT_ORT_L_LL_P_W (OS), IV = QMT_ORT_L_LL_P_W (NOS)

	Top 300							
	Ranker = Eq. 6.1				Ranker = Eq. 6.2			
	Pt = #		Pt = 1.0		Pt = #		Pt = 1.0	
	PP	¬PP	PP	¬PP	PP	¬PP	PP	¬PP
I	**58.64**	54.41	49.87	51.68	51.47	50.56	48.42	49.96
II	**57.78**	54.07	49.55	51.29	41.82	41.82	41.85	41.83
III	**59.76**	55.84	51.67	54.21	52.65	51.75	51.74	52.22
IV	**58.97**	55.60	51.21	53.95	45.34	45.34	45.36	45.35

Table 6.7: I = L_P_W (OS), II = L_P_W (NOS), III = QMT_ORT_L_LL_P_W (OS), IV = QMT_ORT_L_LL_P_W (NOS)

The results show that Type II of feature combinations was most successful making use of: the Top 300 most predictive N-grams for each class, a ranker

that uses Equation 6.1, assigning the N-gram corpus frequency to the class with a matching N-gram in the Top N, and applying a post processing algorithm that sets the highest value to 1 and the rest to 0. To assure validity of our approach, as a sidestep DA classification performances were calculated on the ICSI and Switchboard corpora . The results on the ICSI corpus outperformed the best known results and the results on the Switchboard corpus proved comparable to the best known results (see Verbree et al. (2006)).

6.5.3 Results

For the classification results on the TAS unit labels, as a baseline the share of the most frequent class (Statement) was used (58.92%). Due to the uncertainty of the similarity between the DA classification task and the TAS unit label classification task, we again evaluated the post-processing step, as well as several combinations of features. All the results were obtained after 10 fold cross-validation and are shown in Table 6.8.

FeatureSet	J48	SVM	NB
L-P-W	71.62	72.67	33.18
L-P-W*	71.91	70.53	70.30
QMT-ORT-L-LL-NS	68.08	62.24	54.97
QMT-ORT-L-LL-NS-P-W	73.07	73.20	34.70
QMT-ORT-L-LL-NS-P-W*	**74.41**	72.85	71.73
QMT-ORT-L-LL-NS-LMP-LMW	66.91	66.23	66.85
QMT-ORT-L-LL-NS-RL	67.75	62.24	56.97

Table 6.8: Results on automatic TAS unit labelling. $*$ includes post-processing

The results show that our best result of 74.41 % outperforms the baseline by more than 15%. Two sets of additional experiments were conducted to see whether we can improve the results.

The first experiments made use of language models created with the SRILM toolkit (Stolcke, 2002). We used this package to create language models for each class for both POS (LMP) and Word (LMW) level. This was done because the idea behind the creation of language models resembles the idea of how the P and W feature vectors were constructed. For each test utterance the perplexity values have been calculated using the language models for each class. The lower the perplexity, the more likely the utterance belongs to a particular class (see (Jurafsky and Martin, 2000)). So within the resulting feature vector, the class with the lowest perplexity was set to one, where as the others were set to zero. The results are shown in Table 6.8. The performance turns out worse for the LMP and LMW feature in comparison to the P and W feature. This might have to do with the fact that many data are required to obtain reliable language models.

A second additional experiment was performed to see whether some prede-

fined corpus specific knowledge could help. We exploited the fact that in each of the meetings four predefined roles were played by the participants. Hence it was decided that the role of the participant who uttered the segment was enacted as a feature (RL). One could, for instance, hypothesize that contributions come from specific roles with respect to the classes in which their utterances fall (i.e. the project manager raises more issues than the marketing expert). The results are also displayed in Table 6.8. From the results we conclude that this corpus specific information does not help us here[5].

A combined confusion matrix produced by the SVM classifier on the best performing combination of features is shown in Table 6.9.

a	b	c	d	e	f	$<--$ classified as
3	8	0	20	0	38	a = AIS
3	50	4	58	1	116	b = OIS
2	8	975	883	4	31	c = OTH
9	19	211	3702	32	92	d = STA
0	0	18	166	0	9	e = WST
5	22	6	109	0	298	f = YIS

Table 6.9: Confusion matrix of the SVM-classifier using features QMT-ORT-L-LL-NS-P-W*. AIS = A/B-Issue, OIS = Open-Issue, OTH = Other, STA = Statement, WST = Weak Statement, YIS = Yes/No-Issue.

The table shows that improvements could be made with features that distinguish amongst the classes *statement* and *unknown* and also shows that especially the smaller classes are more often incorrectly classified. Most noteworthy is the 'Weak Statement' class, that is never correctly recognized.

Using ASR instead of Transcripts

To move one more step in the direction of fully automatic node classification, the features were also computed on the ASR data of the AMI meetings following the procedure as described in Ang et al. (2005). The results are shown in Table 6.10. However, one has to be really prudent when interpreting these. The results from the transcriptions cannot directly be compared with the performances on the ASR. This is due to the fact that the class labels originated from the annotations performed on the manual transcripts and the annotations were not conducted on the ASR itself.

From the table it follows that the accuracy drops around 9% when using ASR input for feature extraction, rather than manual transcriptions. This performance drop is comparable to the performance drop that one gets when using ASR data as input for feature extraction for the task of dialogue act classification (see Verbree et al. (2006)).

[5]This finding also seems to contrast Berger's Status Characteristics and Expectation States theory (See Section 5.2.1)

FeatureSet	J48	SVM	NB
L-P-W	65.29	62.34	34.22
L-P-W*	65.14	64.21	64.04
QMT-ORT-L-LL-NS	63.34	58.90	54.39
QMT-ORT-L-LL-NS-P-W	**65.88**	62.34	34.18
QMT-ORT-L-LL-NS-P-W*	65.62	64.14	60.56

Table 6.10: The classification performance using ASR instead of Transcriptions. $*$ indicates preprocessing

6.6 Assigning TAS relation labels to pre-defined relations

In this section we report on experiments related to the automatic labelling of pre-defined relations between speech segments that are labelled with TAS-unit-labels. First an overview of related literature and features is given before our own approaches are explained together with the results that were obtained.

6.6.1 Related work

When focussing on direct text examination in order to detect coherence relations, the word *but*, for instance, can be regarded as evidence of a contrast relation between two adjacent units, *in general* as evidence of a generalization relation and *in other words* as evidence of an elaboration relation. An approach in this direction is explored by Marcu (1997a). His work builds directly upon the recognition of these so-called discourse markers, or cue phrases, such as 'but' and 'in general'.

To detect these cue phrases, one could start from the more general expectation that certain pairs of lexical items are more likely to co-occur with certain discourse relation types than others. A discourse relation r_c hence exists between two utterances or text spans (W_1 and W_2) and is determined by the word pairs in the cartesian product defined over the words in the two spans $(w_i, w_j) \in W_1 X W_2$. Theoretically, any word (pair) from these sentences can signal a particular relation. When using the Bayesian probabilistic framework one could predict the most likely relation r_c by calculating $argmaxP(r_k|W_1, W_2)$ (Marcu and Echihabi, 2002). This approach, however, requires a vast amount of data[6].

Drawbacks of an overreliance on cue phrases as evidence for discourse structure in general and discourse relations in particular is that it makes it difficult to ensure that a computational discourse analyzer will be able to construct a representation that completely covers the text (Corston-Oliver, 1998a). Schauer and Hahn (2001), for instance, showed that in an experiment cue phrases revealed

[6] A zero probability for any of the words in an unseen test sample could disrupt the whole calculation

only 38.8% of the manually labelled coherence relations. In work from Kurohashi and Nagao (1994) besides cue phrases also word and syntactic similarity scores, as well as referential constraints that were derived from cohesive aspects of the discourse and the models into which the text is to be represented were incorporated. To relate segments based on word similarity, lexical chaining techniques were used, whereas for syntactic similarity Part-Of-Speech tags extracted from the segments were compared. Recent work from Reitter (2003) uses cue words, pronouns and punctuation, part of speech categories, lexical similarity and span lengths in order to compose the vector. Corston-Oliver (1998a,b) also incorporated tense and polarity aspects of segments.

Results on relation identification have recently been reported in Wellner et al. (2006). This experiment uses an annotated corpus of news articles, known as the Discourse Graphbank (see Wolf and Gibson (2005)). The corpus contains 8755 coherence relations distributed over twelve different relation types. The identification considered discourse relations between segments *within* the same sentence. An accuracy of 70.04% is reported. Apart from cue phrases and similarity measures the following features were used: the words at the beginning and end of each sentence, the proximity between the segments, the temporal order between the segments, grammatical dependency relations as identified by a sentence tree parser, and so-called event referring expressions[7] that can be temporally ordered.

In an experiment described in Nomoto and Matsumoto (1999) a committee based sampling approach is taken to distinguish segments labelled with two relation types (Elaboration and Sequence) in a corpus with Japanese news articles. The features that they used are the location of the relation in the text, cue words, the previous relation, a similarity measure between the current and the preceding sentence, and a feature specific for Japanese texts related to sentence endings. The approach samples from the corpus to build a number of different models, or committees that each represent a particular label. To classify a test sample, each of the trained models is consulted for their opinion and a voting mechanism is applied for the final label. A performance of 66% is reported on a corpus of 5221 sentences with a baseline of 56%. Recently Murray et al. (2006) investigated the usefulness of prosodic features in classifying five rhetorical relations between utterances in meeting recordings with support vector machines. The results of this study show that with pairwise classification an average accuracy of 68% can be achieved in discerning between relation pairs using only prosodic features, but multi-class classification performing only slightly better than chance (35%).

6.6.2 Features

The features that were used for our experiments are listed in Table 6.11.

The features were split into two sets. The first set contains directly observable features and the second set contains features that require trained language

[7] see Saurí et al. (2006).

Static Features	Speaker Role Source (RS) Speaker Role Target (RT) # words in the source (WS) # words in the target (WT) Depth of node / max depth of complete tree (DT) Depth of node / max depth of current branch (DB) Start time target node / max start time (PE) Source Type (ST) Target Type (TT) Time Difference between source and target in ms. (TM) Word overlap between source and target (OL)
Language Models	Class of lowest perplexity words source (PW1) Class of lowest perplexity words target (PW2) Class of lowest perplexity POS source (PP1) Class of lowest perplexity POS target (PP2)

Table 6.11: Features that were collected for the classification of TAS relation labels.

models in order to be computed. The word overlap feature was computed by using twice the subset divided by the total amount of words so that a maximum value of 1 was assured.

We used the SRILM package (see Section 6.5) to calculate the language models. We did not pursue the N-gramming method nor did we calculate $argmaxP(r_k|W_1, W_2)$, as initial experiments failed miserably. One reason for this could be the unequal distribution of the class labels, or the fact that all text spans are incorporated in at least two different relation classes, (once as source and once as target). In an attempt to overcome this drawback, it was decided to compute the language models for the source and target nodes for each of the relation types individually. Perl scripts were again used to obtain the feature values from the corpus data.

6.6.3 Results

As baseline for the experiments, we again used the share of the most frequent class label. In this case 'positive' with a share of 47,36%. All the results are shown in Table 6.12.

From the table it is shown that the language model features are still performing very badly in comparison to the other (static) features. A possible hypothesis for this, that could be worth exploring, is that the features work very well on the training data but that they do not generalize on the test data. This results, due to the higher weights of the features obtained from the training data, in a drop of the final performance. This phenomena, known as overfitting, occurs even when other features were present. It indeed appears that the

FeatureSet	J48	SMO	NB
Language Models and Static Features	41.02	47.80	49.21
PW1-PW2-PP1-PP2 (I)	36.12	35.87	32.70
RS-RT-WS-WT-DT-DB-PE-ST-TT-OL-TM (II)	54.52	**57.09**	55.64
WT-ST-TT*	56.50	56.67	**57.03**

Table 6.12: $*$ = subset with most predictive features, I = Language Models, II = Static Features

performance increases when just the static features are used. A multi-class performance of 57% is comparable to that of Marcu and Echihabi (2002). Feature reduction resulted in a most predictive subset that contained the features WT, ST and TT. The performance of these three features turns out to be nearly as good as the performance obtained by using all static features (see Table 6.12).

When we take a closer look at the distribution of one of the best predicting features, we obtain for the WT feature (number of words in the target) Figure 6.3. From the figure it shows that differences exist between the average feature value for each class. These value differences have been exploited by the classifier.

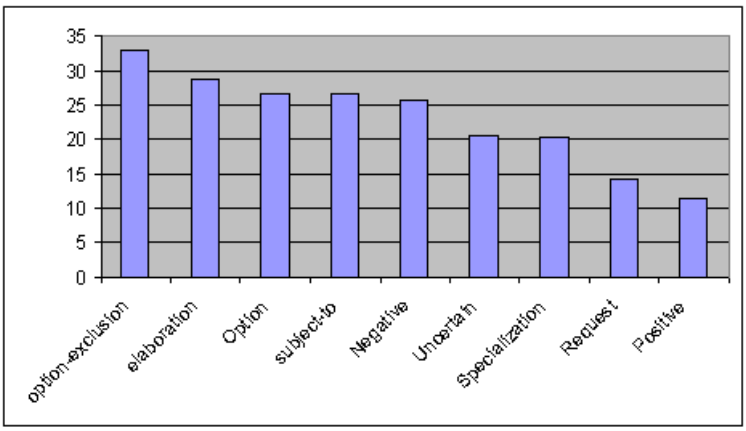


Figure 6.3: An overview of the average number of target words per relation category.

The force from this feature can be made explicit by looking at its value differences of the various classes. A significance test shows, for instance, that 18 of the 38 possible pairwise combinations of class labels are significant at a $P < 0.05$ level and that 25 of the 36 possible combinations are significant when using $P < 0.1$. When looking at a less discriminative feature, PE for example, only 8 of the 36 pairwise label combinations have significant different feature values for $P < 0.05$ and 7 of the 36 for $P < 0.1$. Another way to compare

the strength of the feature is to look at the performance of the classifier when omitting the features from the classifier input. When using the NB classifier and all static features (except the feature under examination) the performance dropped with 0.96% when the WT feature was omitted and the performance increased (!) with 0.19% when the PE feature was left out.

6.6.4 Adding Templates

To see whether common sense cue phrases could aid us, a number of templates that were expected to be associated with the class labels was devised. The templates contain words that were expected to be relevant in the source and/or the target. This list is shown in Table 6.13.

Relation	Template
Positive	-I agree -yeah -okay -right -sure -alright -true
Negative	-no -I don't -I don't agree
Uncertain	-mm-hmm -maybe
Option	-how about -it could be -I think -I like -I might -we could -I would -we should probably
Option exclusion	-let's abandon -I would not
Elaboration	-I mean -you know
Specialization	-more specific -should we say -I would say
Subject-to	-depends on -apart from -keep in mind -limited to
Request	how about- what about- what do you- how do you-

Table 6.13: '-' Separates Source and Target

A binary presence value was added to the feature vector for each of the templates. The performance for the templates, and for the templates in combination with some other features is shown in Table 6.14.

FeatureSet	J48	SVM	NB
Templates	48.99	47.85	46.99
Templates + Static features	55.63	57.53	55.61
WT-ST-TT & -yeah *	**58.14**	57.30	57.29

Table 6.14: Performance results from the templates and the templates combined with the all other features. $*$=most predictive subset.

Feature reduction on the set of templates and static features revealed that the '-yeah' template (the word yeah occurs in the target node of a relation) in combination with the other predictive features improves the performance with around 1%. The fact that the templates in combination with the static features result in a lower performance than before can again be attributed to the fact of overfitting.

6.6.5 Binary Classification

To make our results comparable to those mentioned in Marcu and Echihabi (2002), we also computed the pairwise classification results (binary classification) with our best feature set (WT,ST,TT,-yeah). The work from Marcu and Echihabi (2002) showed that binary discourse relation classification can be done relatively well when the models are trained on extremely large data sets. We pursue this line of research by calculating the performance on paired relations, but now with smaller data sets. All sets contain an equal number of samples for both relations, which results in total to twice the number of data samples for the class with the smallest number (see Table 6.2). The baseline for each of these cases is 50%.

Results

Table 6.15 lists the results of the binary classification experiment for the J48 and the SVM classifier.

		Spec.	Pos.	Opt.	Req.	Unc.	O.-E.	S.-to	Neg.
SVM	Elab.	78.24	77.87	78.62	93.72	70.27	85.71	69.47	65.60
	Spec.	-	79.39	74.05	82.06	67.18	71.42	62.21	75.95
	Pos.	-	-	83.86	97.76	76.01	64.29	79.21	78.13
	Opt.	-	-	-	93.95	74.32	46.42	69.47	78.34
	Req.	-	-	-	-	80.94	92.86	84.74	95.52
	Unc.	-	-	-	-	-	89.29	57.89	69.05
	O.-e.	-	-	-	-	-	-	46.42	78.57
	S.-to	-	-	-	-	-	-	-	67.11
J48	Elab.	69.85	78.16	78.12	92.83	65.06	53.57	67.37	67.94
	Spec.	-	81.30	74.43	77.48	60.69	82.14	56.49	72.14
	Pos.	-	-	82.20	97.09	71.37	75.00	81.32	79.09
	Opt.	-	-	-	92.60	70.46	64.29	65.79	78.13
	Req.	-	-	-	-	75.78	96.43	78.95	95.52
	Unc.	-	-	-	-	-	60.71	61.05	65.76
	O.-e.	-	-	-	-	-	-	42.86	46.43
	S.-to	-	-	-	-	-	-	-	63.95

Table 6.15: Matrix of the performance results for binary classification using SVM and J48

It should be noted that the number of samples that was used to compute the results for the option-exclusion relation (28) is very low. The results obtained for this label therefore do not provide more than an indication as to whether or not it can be successfully automatically distinguished from others, rather than hard evidence. The results outclass those mentioned in Marcu and Echihabi (2002). From the table it further shows that especially the 'Request' and the 'Positive' relation distinguish relatively well from the others.

6.7 Application

Eventually the possible applications for meetings annotated with the TAS schema are endless. The resulting trees can be used by other applications as a source for information retrieval purposes, or to aid question answering systems in the context of a meeting browser. They can be used for automatic summarization purposes, and for processes that aim to find out who adhered to a specific opinion at any given moment. Managers can, for example, use the diagrams to investigate what went well or wrong in the discussion, which arguments were made in favor of or against a specific proposal, who proposed the accepted solution, or who objected to most of the discussed points. RST has, for instance, already been used in the context of summarization (Corston-Oliver, 1998a), and machine translation (Marcu et al., 2000). Burstein et al. (2003) used automatic text structuring in the context of learning students to write coherent essays, as they often have difficulties in structuring their sentences.

For an end user it is said that argument diagrams themselves provide a representation leading to quicker cognitive comprehension, deeper understanding and an enhanced detection of weaknesses (Schum and Martin, 1982; Kanselaar et al., 2003). Furthermore they are said to aid the decision making process, and can be used as an interface for communication to maintain focus, prevent redundant information and to save time (Yoshimi, 2004; Veerman, 2000).

Although one way to evaluate the trees is to compare them with manually created trees, another option is to evaluate the impact of the tree on tasks they will be used for. We therefore devised a test to measure the usability of the argument diagrams themselves in comparison with other representations of the same discussion.

Method

Stimuli: Imitating a user that wants to ask a question to a browser system, we created a list of hard to answer multiple-choice questions about the contents of six similar discussions about the design of a remote control. These questions were shown on a screen to subjects using a newly created software package. The answers could be found in provided representations of the discussions printed on a piece of paper. Each question could be answered by selecting the answer with the mouse.

Procedure: We provided the discussions in one of the following three representations: (1) a printout of the raw transcriptions of the discussion, (2) the transcription with a colored background in correspondence with the labelled unit segments of the TAS-schema and (3) a TAS-argument diagram.

Before the start of the experiment the subjects were asked to read a document describing how to read and interpret the representation of the discussions presented. Next, each subject was asked to complete as many questions as they could answer about 6 discussions. It was impossible to answer all questions within the given time frame and the same questions were asked in the same order in each of the conditions. After exactly five minutes the subjects were

asked to proceed with the next discussion. Apart from giving the answer we also asked for the perceived difficulty of the question and measured the time it took to complete each question. No breaks were allowed in between the discussions and the subjects were asked not to start reading the discussion before the first question was shown on the screen. Subjects did not receive any feedback on their judgements.

Participants: A total of 30 persons (25 male and 5 female) participated in the experiment. Resulting in 10 completed experiments per condition or representation. The participants were students and employees of our department in the age range between 22 and 59.

Results

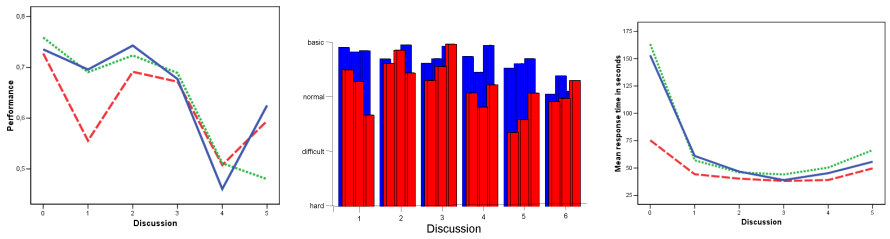
Since the number of participants does not allow us to make hard inferences, our findings should be regarded rather as indicative. The most important results of our experiments are shown in Figure 6.4. In total 887 questions were answered, from which 567 were correct. Figure 6.4(a) shows the performance (percentage correctly answered questions) of the subject in each of the conditions. The continuous (blue) line corresponds to the subjects using the argument diagrams, the short dashed (green) line for the colored transcripts, and the long dashed (red) line for the raw transcription.

The figure shows that for the first four discussions the performance of the raw transcript is lower than for the other two conditions. This possibly indicates that the extra information embedded in the unit labels, which is contained in both the other two conditions, could result in the observed performance increase. No significant differences were found with respect to the total number of correctly answered questions.

When looking at the response time for the correctly answered questions, it seems that for the first two discussions the subjects need to get used to the unknown representation types. For all the other discussions it appears that the raw transcription condition results in a quicker answer, although the difference for the last four discussions is much smaller than the difference for the first two discussions. It is interesting to note that the time required for finding the answers with the argument diagram for the last two discussions is shorter than for the colored transcripts.

The findings for the perceived difficulty are depicted in Figure 6.4(b). The bars at the back correspond to correctly answered questions, whereas the bars at the front correspond to wrongly answered questions. It appears that when the questions were answered correctly, in four out of six discussions the questions were perceived as more easy when provided with an argumentation diagram (third column), than when provided with another representation (first column = raw transcript, second column = colored transcript). It should be noted, especially for the first discussion, that not only the perceived difficulty is harder, but also the the response time was much longer.

Though it can be that the differences are caused by differences in the questions and the individual discussions, it appears that people need time to get



(a) Performance (b) Difficulty (c) Response time

Figure 6.4: Some results of the user experiment

used to the argument diagrams in order to reap their benefits. Once used to the method people perceive the questions as less hard and the extra information embedded in the unit labels seems to increase the performance.

6.7.1 JFerret implementation

A plug-in has been developed for the JFerret meeting browser (Wellner et al., 2004). This plug-in, shown in Figure 6.5, enables users to access the discussions depicted on a meeting time line. For each discussion the resulting argument diagram appears and thereby allows a quick grasp of the content of the on-going discussion. Clicking on the nodes in the diagram shifts the browser directly towards the corresponding moment in the meeting.

The current implementation of the plug-in works with the manual annotations and can be accessed by other plug-ins.

6.8 Final Thoughts

This chapter introduced a method to capture argumentative aspects of meeting discussions in a way that an argument diagram can be created that shows how the discussion evolved. A corpus containing over 250 argument diagrams derived from real-meeting discussions has been created and machine learning experiments for automatic labelling resulted in a performance of 74.41% for pre-selected units and 58.14% for pre-selected relations. The obtained performances are comparable to performances mentioned for similar tasks, but in different domains.

Remarks that were made in the direction of the TAS annotation scheme regarded the fact that we opted for a tree structure instead of a graph structure (see e.g. Pallotta et al. (2006)). I am still confident that this was a good choice. Arguments have been made about the readability, as well as the fringe on which new contributions can be attached. On the other hand it is true that old issues that are part of a different branch can be re-addressed and that TAS in this case does not contain explicit links between them. I assume, however, that on top

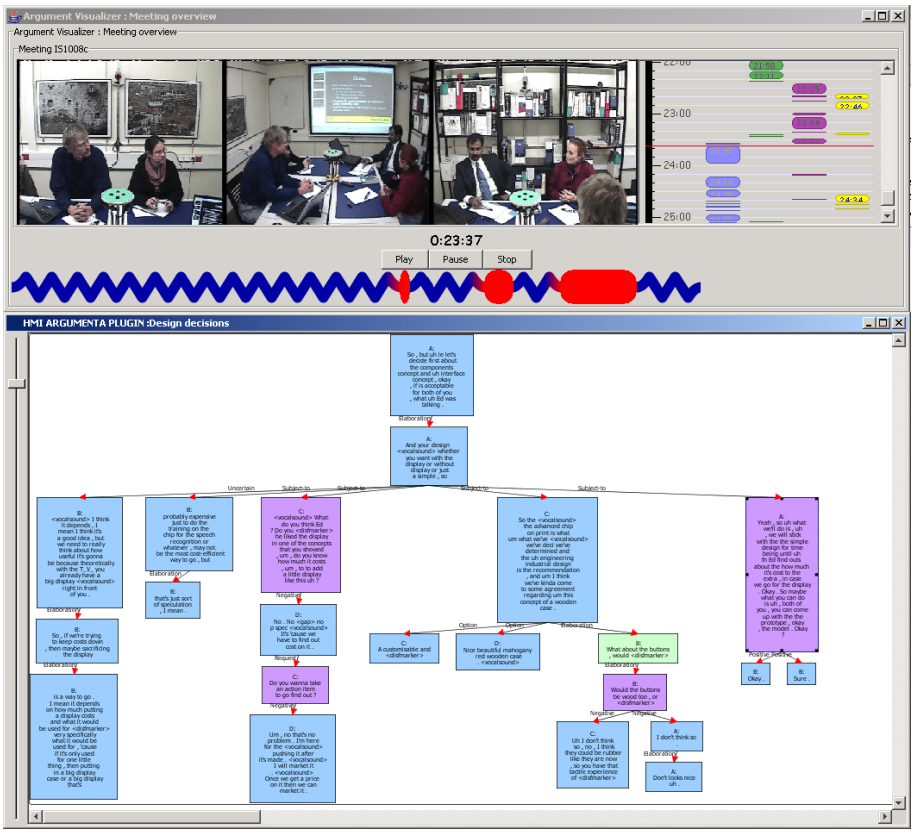


Figure 6.5: Argument-Diagrams as a JFerret browser plugin

of the current TAS trees new algorithms can be created that detect themselves that a certain issue is re-addressed. In case the goal is summarization, one could for instance, rather than showing how the discussion enfolded in time, show a derivative that contains the information that was discussed per topic that was addressed.

For the unit labelling the work has mostly concentrated on N-grams of words and POS-tags. Results shown in Tables 6.6 and 6.7 indicate that the N-gram-selection methods have had quite some influence on the performance. I had to limit the number of variables, as well as the possible values of the variables (mentioned in Section 6.5.2) due to the rapidly increasing number of experiments. More research on scoring algorithms might perhaps yield better N-gram selection methods and a better performance. Also, the points attributed to a feature when an N-gram is present can be re-considered, just as the number of cueing N-grams that are used. Here I would also like to mention that the use of the presence or absence of a question mark as feature, de facto runs

ahead of the current state of technology, as in automatic speech recognition it is very hard to recognize whether an utterance is a question or not (see e.g. Kim (2001); Huang and Zweig (2002)). The last thing I would like to remark is that we considered the problem as a multi-class classification problem. Possibly, a stepwise approach that distinguishes one label at a time, by following a binary classification approach, will be useful (see e.g. Fernandez and Picard (2002) for such an approach.).

The relation labelling attempt was, although comparable to other approaches, with respect to the obtained performance, less successful than the performance obtained for unit labelling. Possible reasons for this are in the first place the number of relation labels. The more labels, the harder it is for a classifier to assign the right class. In the second place, there is the inter-annotator agreement. The lower the inter annotator agreement, the lower one can expect the performance of the classifier to be[8]. The Virtual Kappa values for the relation labels show in Table 6.4 that the maximum score for training and testing on a different annotator is 62.53%. A machine learning performance higher than this is, from my point of view, impossible.

The road towards full argument diagram creation is still long and the attempts described in this chapter confined themselves to the two classification steps on this road. The segmentation steps of segmenting the meeting into discussions and segmenting the utterances into units that can be labelled are both challenging tasks. In the area of dialogue act recognition current approaches start to combine the segmentation and the labelling task (cf. Zimmerman et al. (2006)). Not to forget is the relational finding. Although it was proven that most units connect to units that directly follow each other in time, the argument structure, in essence, is shaped by the relations that do not connect units that directly follow each other in time. This makes the challenge perhaps even greater.

[8] Jovanovic et al. (2006) on the contrary report an accuracy of nearly 80% on an addressee classification task with a reported κ for the class labels between 0.45 and 0.67.

Chapter 7

Relating Influence and Argumentation

7.1 Introduction

When talking to others about my PhD work, I have often been confronted with remarks about the 'connection' that should exist between the phenomena of argumentation and influence. The expected linkage between the two higher level phenomena debated in the last two chapters seems indeed logical and also interesting to investigate. Once one uses valid arguments, one can change the attitude of the other, and a change in attitude can in turn be regarded as a sign of influence. This change of attitude, however, is not by definition realized by just using valid argumentation, nor is being influential a necessary requirement for putting forward valid argumentation.

Amongst the several definitions that exist for argumentation Van Eemeren et al. (1987) define argumentation as a social, intellectual, verbal activity that serves to justify or to refute an opinion, consisting of a constellation of statements and that is directed towards obtaining the *approbation* of an audience. The interesting word here is approbation which is closely related with the ability to get something approved, regardless of its truth value. Through approbation, a change of attitude, or belief towards a certain issue, is to be established. So, independent of the fact whether the message you are trying to convey is true by itself, one could still bring arguments trying to *persuade* the other. Persuasion in turn, relates to influence, as it guides people towards the adoption of an idea, a claim, an attitude, or an action.

Persuasion is highly related to what in ancient Greece was called rhetoric. Rhetoric is the art of good and cogent oratory and is concerned with '*how* a message is brought, rather than *what* message is brought. Three factors are generally distinguished: ethos (how the character of a person influences the audience to consider him to be believable), pathos (how emotions affect the audience by using sentiment, joy, sorrow, love or hate) and logos (how the use

of language affects the message or logical means of convincing). A clear relation here emerges with the dominance and influence theories described in Chapter 5. Interestingly, the concept of ethos seems to align with the Status Expectation Theory of Berger et al. (1980), whilst the concept of logos seems to be more in line with the theory of Lee and Ofsche (1981), stating that the interaction itself is responsible for the establishment of status differences. So given these observations it is not unlikely to expect that the two phenomena indeed are somehow related.

The aim of this chapter is to test the hypothesis that these phenomena indeed are related on the basis of some empirical grounds. Given the data sources that were collected for the experiments in the two previous chapters, Section 7.2.1 investigates statistical dependencies and (cor)relations between the TAS-units and the influence levels that were obtained from the questionnaires described in Section 5.5. It is examined whether, at the meeting level, differences exist amongst the various dialogue act distributions for the various influence types also if certain TAS relation types co-occur more frequently with participants of particular a influence level. Section 7.3 of this chapter then examines whether the classification performance, as described in the previous chapters, of the phenomena can be improved when using one phenomenon as a feature of the other and vice versa. Section 7.4 then examines whether certain regularities can be mined in an unsupervised manner from the combined data sets, before Section 7.5 tries to provide a profile of 'High' influential participants in relation to 'Low' influential participants by combining the results of the first three sections.

7.2 Statistical explorations

This section reports on experiments that were conducted to find out whether there are any dependencies between items of the TAS scheme and the participants influence rankings. These dependencies could provide valuable insights into how the phenomena described in the previous chapters manifest themselves and whether perhaps new valuable features exist that can be used in future classification systems.

7.2.1 TAS units in relation to influence levels

The exploration was started by conducting three different kinds of experiments to see whether, and if so which, aspects in relation to the TAS unit labels could be (cor)related to the various influence levels. Examined for possible relationship with the influence rankings were: the total number of units, the average unit duration, and the unit type distributions. The experiments are reported below.

After merging the data sets that were used in the previous chapters, influence information (as acquired from the participants themselves) was available for 29 discussions distributed over 18 meetings and thereby covering 865 of the total of 6920 TAS unit labels. 263 of these TAS units were uttered by a 'High' influential participant, 474 by a 'Normal' influential participant and 155 by

a 'Low' influential participant. The distribution of the unit labels comprised 4 A/B Issues, 24 Open Issues, 51 Yes/No Issues, 464 Statements, 20 Weak Statements and 302 Others. To increase the number of samples per category, the argument labels were grouped into three main categories (Issues, Statements and Other). All in all it resulted in the data set that is shown in Table 7.1.

	low	Normal	High	Total
Issues	12	40	27	79
Statements	78	254	152	484
Other	65	153	84	302
Total	155	447	263	865

Table 7.1: Distribution of label combinations for combined argumentation (merged) and influence data.

A first exploration reveals that the distribution of argument labels as a function of the influence values does not turn out to be significant ($\chi^2(4, N = 864) = 4.73, P < 0.31$), nor do ANOVA tests on the individual labels show any significant results. As a consequence, one might conclude that both phenomena seem to be independent.

Not taken aback by this rather 'disappointing' result some closer looks were taken to investigate for possible other interdependencies.

Examining the number of TAS units When considering the number of TAS units uttered per person per meeting, an average of 7.27 was found with a standard deviation of 3.56. No significant differences were found with respect to the number of turns for each type of influence level. When zooming in on the contribution of turns along the discussion (split up in five bins of equal time intervals) we obtain Figure 7.1.

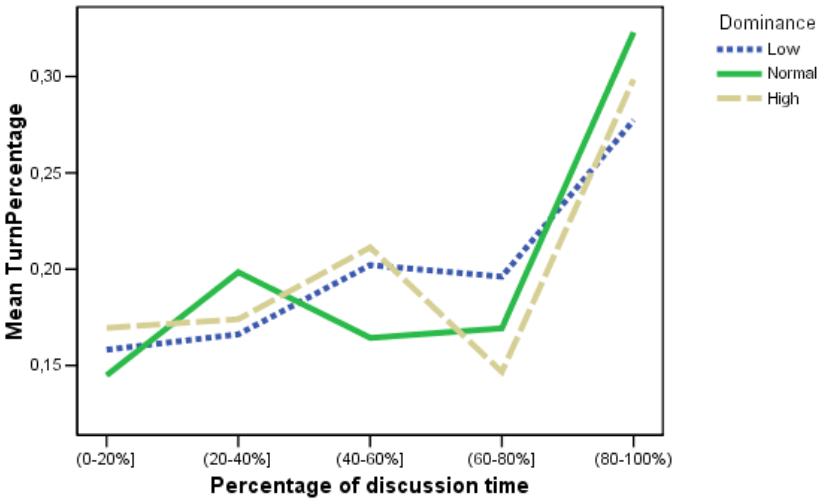


Figure 7.1: The fraction contributions divided over five time intervals per influence type.

Apart from the fact that no difference exists in the total number of TAS units uttered per influence level, no significant difference for the various influence levels when considering the number of TAS units uttered per bin were found. A significant positive correlation, however, was found between the fraction of turns and the progress of the discussion for all influence levels combined (Pearson's correlation coefficient r=0.22, with a significant regression model $F(1) = 30.34, P < 0.001$) as well as for the separate influence levels. (r between 0.24 and 0.19, $P < 0.01$ for 'Medium' and $P < 0.03$ for 'Low' and 'High').

This finding shows that towards the end of the discussion people tend to talk in shorter turns. A logical explanation for this might be that people reach agreement towards the end, and that contributions are in terms of 'yeah' and 'sure' occur more frequently. Another, but, from my point of view, less likely explanation could be that people start to run out of time and therefore try to limit the length of their contributions.

Examining the duration of the TAS units To examine in more detail the finding that turns towards the end of a discussion seem to be a little shorter, the average duration of the turns was computed for the same discussion intervals. The results are shown in Figure 7.2.

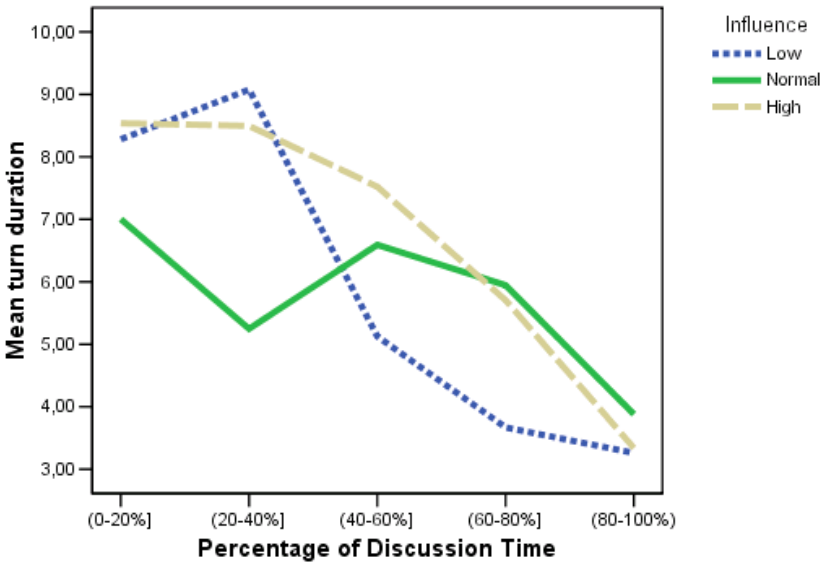


Figure 7.2: The average turn duration over a discussion per influence type.

Statistical analysis on this data revealed a significant decrease in turn duration as the meeting progresses for all the influence levels individually (Pearson's correlation coefficient r between -0.18 and -0.11, $P < 0.01$ for 'High' and $P < 0.03$ for 'Normal' and 'Low') as well as for all levels combined (r=0.15 with a significant regression model F(1)=19.66, $P < 0.001$). This was expected, when looking at our earlier finding that the number of turns increases along with the meeting. One could, when considering Figure 7.2, get the idea that less influential people generally resort to shorter turns more quickly than more influential people. This is interesting, because most decisions are taken at the final stage of a discussion. However, when considering the individual time intervals, one-way ANOVA with post-hoc Fisher-LSD testing showed no significant differences between the average duration of the turns for the various influence levels. This again could be so due to our relatively scarce data set.

7.2.2 Dialogue acts and influence

It was shown in the previous section that for the three grouped argument labels no significant differences existed in their distributions over the three influence categories. A possible explanation for this could be the relatively few examples. This, combined with the fact that we grouped the unit labels, was the reason to conduct some extra experiments. It was decided to examine more closely whether and how, certain categories of dialogue-acts can be related to the various influence rankings over the course of a meeting. It must be said that the dialogue

acts, in line with the TAS units, themselves can be regarded as a higher level meeting phenomenon. This implies that they are not directly observable, and thus that they were not included in any of the feature sets described in Chapter 5. But as they are available for 30 of the 40 meetings for which influence rankings were available (see also Section 6.5.1), I could not resist the temptation to at least explore the data for some interesting findings.

As a first attempt the fractions for the occurrence of all dialogue-acts was computed for all participants. These results were subsequently merged for each of the influence levels. The resulting average fractions are shown in Figure 7.3.

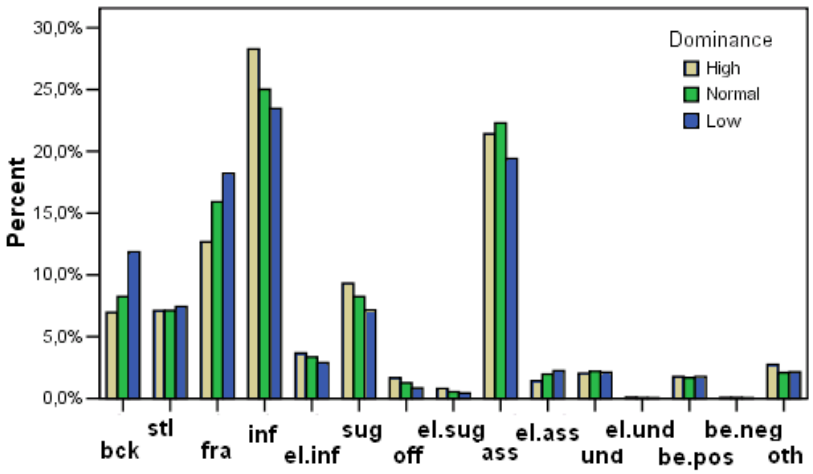


Figure 7.3: The fraction of dialogue acts per influence level.

From Figure 7.3 is seen that there seem to be some interesting differences between the various dialogue act distributions. Statistical analysis by means of ANOVA showed that on the $P<0.05$ level significant differences exist for the labels 'fragment' ($F(2)=7.87$, $P<0.001$), 'backchannel' ($F(2)=6.01$, $P<0.003$), 'elicit-suggestion' ($F(2)=3.94$, $P<0.022$) and 'suggestion' ($F(2)=3.19$, $P<0.045$).

For all of these some intuitive explanations can be given. Starting with the 'fragment' label, it appears that people who are highly influential utter less fragments than people who have low influence. This finding is in line with the finding from Bales et al. (1951) discussed in Section 5.2.1 who stated that people who are interrupted more than others are likely to be of a lower social status, and hence likely to be less influential. For the 'Backchannel' label it appeared that people who are 'Low' on dominance backchannel more than people who are 'high' on dominance. One could say that those that backchannel signal to others that they follow, or that they express listeners' behavior (Yngve, 1970). By providing backchannels people signal that they understand the messages submitted by others. One could therefore say that it can be related

to their *participation* level and hence to their talkativeness. This aligns with Bales (1950) who observed that people who talk more than others are likely to be more dominant. Both of these dialogue-act labels are related to the meeting process. The remaining two labels 'suggest' and the 'elicit-suggest' show that both types of utterances are uttered relatively more by people who are 'High' on dominance than by people who are 'Low' on dominance. Both the elicitation of suggestions, as well as making suggestions during a meeting, or a discussion, relate to the fact that people provide options, or ideas, that could be solutions to the problems, or issues at hand. This finding, hence seems to provide evidence for the hypothesis that dominance and argumentation are related. An interesting aspect is that these two labels appear more related to the task than to the process. This observation aligns with the fact that the annotators were asked to rank the participants (see Section 5.2.2), without mentioning the various dimensions of a meeting, as described in Section 2.3.2. On the other hand this finding also suggests that the phenomenon manifests itself in more than one meeting dimension.

The data was again transformed into a feature set in a similar manner as described in the previous section. For this experiment it resulted in a data set containing 120 samples, out of which 25 were labelled 'High', 69 were labelled 'Normal' and 26 were labelled 'Low'. The results are shown in Table 7.2.

FeatureSet	J48	SVM	NB
All Dialogue-acts	56.66	58.33	45
Fragment and Suggest*	55.83	57.5	53.3

Table 7.2: Results on automatic influence level classification using the fraction of dialogue act labels as features. '*' = best subset.

Given the majority class baseline of 57.5% it appears that, although some of the feature values differ significantly, the features themselves are unable to outperform the baseline. Also after applying a post-hoc feature analysis this turned out to be impossible[1].

7.2.3 TAS Relations and Influence

This Section reports on attempts to relate the various relations that exist between nodes in the argument diagrams to the levels of influence. Similar to the previous Sections, for each participant, for each meeting, the percentage of relation labels was sampled. The combined data resulted in a data-set of 59 participants, participating in 15 meetings (not in all meetings were discussions, nor did all participants participate in all discussions). 13 of the participants were labelled as 'High', 33 of them were labelled as 'Normal' and 13 of them were labelled as 'Low'.

[1] Note that the optimal feature set contains the 'fragment' and 'suggest' labels which, given the significance levels and their complementarity in distinctiveness (see Figure 7.3), is a logical choice.

An overview of the 95% confidence interval of the mean percentage of the six most frequently occurring relation labels is shown in Figure 7.4.

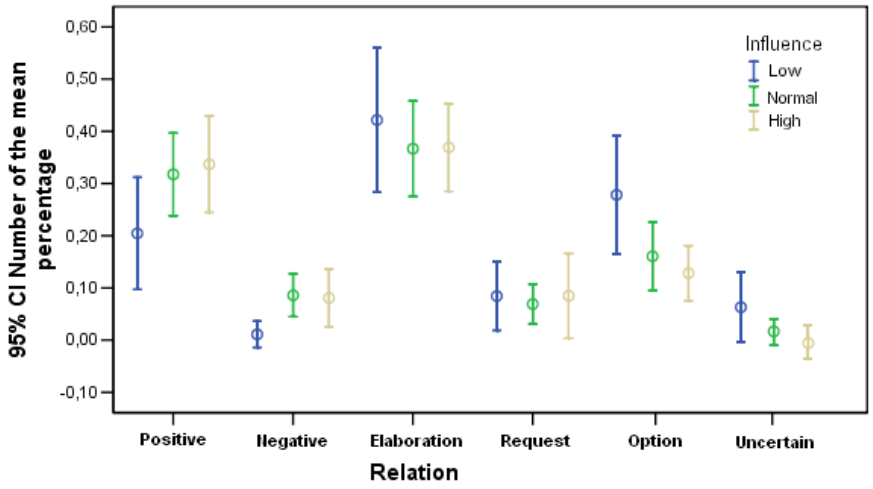


Figure 7.4: The mean number of relation occurrences per influence level.

ANOVA testing showed a significant dependency between the 'uncertain' relation category and the influence levels ($F(2)=3.52$, $p<0.037$). It appears that the lower the participant's influence, the more uncertain, or unclear, his or her contributions to the discussion are. Spearman's correlation coefficient ρ however did not prove significant. The relatively low number of samples plays us parts here. For all the other relations we therefore cannot draw any hard conclusions.

When considering Figure 7.4 one could, however, construct the hypothesis that evaluative contributions, in terms of 'positive' and 'negative', seem to occur more frequently for higher influential participants. So if you give your opinion on things you might become more influential. But again, this is just a tendency that can be observed from the figure and this is not based on significant evidence.

Another interesting observation that can be made is that it seems that people of low influence seem to provide more 'options' to the discussions. These options relate to possible answers to issues that were raised. This is quite remarkable, because on the one hand this is perfectly in line with Wang (2006) who stated that the more dominant participants ask the questions. But on the other hand, it seems to contradict the statistically significant finding from the previous section saying that suggestions are mostly put forward by influential participants. However, one should note here that the dialogue acts that were considered go beyond the discussion boundaries and that the suggestion label, as is formulated in the annotation manual is applied in relation to "when the speaker expresses an intention relating to the *actions* of another individual, the

group as a whole, or the group in a wider environment". This shows that the suggestion label covers more than suggestions for solutions to particular issues (options) and also that a greater 'force' lies behind it in a sense that it really steers towards an action, rather than just raising an idea.

7.3 Cross-fertilizing features

A typical question we want to answer from a machine learning point of view, for example, when considering Figure 7.4 deals with the extent to which the different distributions of certain (class)labels are useful for the classification process. Even more, since the previous section also showed that indeed some regularities seem to exist between the level of influence of a participant and the way that argumentation enfolds in a discussion. This section therefore aims to investigate the usefulness of (the features of) one of the phenomena of influence and argumentation as predictor, or feature, for the other phenomenon in a machine learning context.

One should note that both phenomena are higher level phenomena and that, in a real life situation, it is not a clever choice to predict higher level concepts with other higher level concepts. This is true for at least two reasons. In the first place, the recognition process of the phenomenon that serves as a feature has to be recognized itself. This in turn requires more and other directly observable features. In the second place is it quite unlikely that the recognized higher level concepts are free of errors. The aim of this section is therefore just to investigate the extent to which one phenomenon theoretically could improve the recognition of the other and to explore the consequences of interrelation for the classification performance.

Theoretically, one could expect that whenever a certain feature f aids a classification task C, a classifier would be better able to distinguish the class labels and hence the recognition rate would increase. If this feature, however, represents a class label that itself can also be recognized by a different feature set $\{f_1, ..., f_n\}$ one could choose to replace f with the set of features that was devised to recognize f itself. This is interesting because one could expect that the recognition performance of C will be influenced by the fact that the function from $\{f_1, ..., f_n\}$ to f is not error free, and hence it could be that the performance of C will be higher when using manually assigned values for f, rather than a whole set of automatically obtained features that are only to a certain extent able to represent f. However, as the feature set contains more than one feature, it could also perfectly well be that a certain feature of f (e.g. f_3) is more beneficial to C than f itself. For this experiment we confined ourselves to the manually assigned class labels that were elicited from the meeting participants and the manual annotations.

7.3.1 Predicting influence with argumentation

The first experiment tries to predict the influence level (dependent variable) making use of just the argumentation label distributions (independent variables).

As influence was measured on a meeting level, the feature vectors were also created on a meeting level by taking the label fraction distributions for the individual participants as feature values to predict the influence label of the associated participant. This resulted in 59 samples[2] with a baseline of 55.93%. Machine learning algorithms were trained and evaluated using 10-fold cross validation. The results are shown in Table 7.3.

FeatureSet	J48	SVM	NB
STA-WST-OTH-OIS-AIS-YIS (unbalanced)	55.93	55.93	54.24
STA-WST-OTH-OIS-AIS-YIS (balanced)	25.64	25.64	25.64
STA-OTH-ISS (unbalanced)	54.23	55.93	52.54

Table 7.3: Results on automatic influence level classification using the fraction of argument labels as features.

From Table 7.3 it appears that on the balanced corpus none of the tested classifiers outperforms the baseline. Not with the class labels added as feature, nor with the features that predict the class label, nor after merging the different issues and the different statements.

To explore this finding, a multiple linear regression model was instantiated from the data. It not surprisingly appeared that none of the coefficients proved significant, nor for the individual labels, nor after merging the statements and the issues (the stronger the correlation coefficients, the more discriminating the feature).

7.3.2 Predicting argumentation with influence

For the second experiment the influence labels were used to see whether they could aid the prediction of TAS labels (both units and relations). So in this case the class labels were the TAS labels and the influence value of the speaker was added as a feature. The results are shown in Table 7.4.

The results indicate that the dominance feature does not seem to be of any use to the classifier. For the nodes of the TAS schema, the dominance feature itself does not score above the baseline of 55.95% (most frequent class is statement(464) amongst a total of 865 labels.). When adding the dominance feature to a set of more useful features (see Section 6.5.3), the performance does not increase either. For the relations of the TAS schema the baseline is set by the elaboration relation (181 occurrences amongst a total of 525 relations) to 34.4%. Again here the dominance feature does not prove useful, neither in combination with a set of other features that are useful (see Section 6.6.3).

[2] 13 were labelled as 'High', 33 as 'Normal', and 13 as 'Low'.

Class	Feature set	J48	SVM	NB
Nodes	DOM	53.64	53.64	53.64
	QMT-ORT-L-LL-NS	73.53	68.21	64.05
	QMT-ORT-L-LL-NS-DOM	71.91	68.32	64.05
Relations	DOM	34.48	34.48	34.48
	TT	39.24	39.24	39.62
	TT-DOM	38.86	39.43	38.29
	TT-WT	44.95	39.80	44.00
	TT-WT-DOM	43.62	42.67	44.57

Table 7.4: Results on automatic TAS unit labelling with and without the dominance (DOM) feature

7.4 Rule Induction

The findings from the previous section suggest that although some link between influence and argumentation exists in the data, the existing differences cannot be exploited for classification purposes. This, however, does not withhold us from further digging into the data in order to look for some other dependencies. Section 7.2 already examined the combined data set in a statistically supervised manner. This section will use an unsupervised manner known as association rule mining to explore the data. Association rule mining finds associations and/or correlation relationships among large sets of data items. These resulting association rules bring to light feature value conditions that co-occur in any given data set. Association rules contain a precondition (antecedent) and a conclusion (consequent). The precondition is a series of constraints that is laid over the features and the conclusion generally gives the label that applies to instances covered by the constraints. An association rule can typically be expressed by an 'If a Then b' clause, where the preconditions are specified in the a part and the conclusions in the b part (Witten and Frank, 2000).

Because many different association rules can be derived from even a tiny data set, interest is restricted to rules that apply to a reasonably large number of instances and have a reasonably high accuracy on the instances they apply to. The resulting rules that are found are therefore usually ranked according to their 'strength'.

The 'Tertius' algorithm (Flach and Lachiche, 2002) that was used in our case for rule mining presents two measures for the strength of the rule: the confirmation value[3] and the frequency of counter-instances (the number of counter-instances divided by the total number of data items). A rule is said to be better than another if it has a higher confirmation value. Another rule mining technique, which we did not use, is called Apriori (Agrawal et al., 1993). Apriori

[3]The confirmation value trades off the decrease in counter-instances from expected to observed and the ratio of expected but non observed counter-instances (see Flach and Lachiche (2002) for more detail).

uses support (the fraction of the data set on which the rule can be applied) and its confidence -or accuracy- as evaluation metrics.

A first experiment was restricted to the combined data set, as described in Section 7.2.3, containing both the influence class labels and the grouped argumentation node and relations labels on the meeting level. Considered are the label values of the argumentation node fractions in relation to the influence values. The node fractions were discretized in three nominal categories 'High', 'Normal' and 'Low' using WEKA's simple binning algorithm Witten and Frank (2000), so that Tertius rule induction could be performed. The top three rules are shown per influence category in Table 7.5.

I	II	Antecedent	Consequent
0,164	0,448	Sta = 'normal'	LOW
0,155	0,405	Iss = 'low' and Sta = 'normal'	LOW
0,103	0,155	Sta = 'normal' and Oth = 'high'	LOW
0,145	0,000	Iss = 'low' and Sta = 'low'	NORMAL
0,112	0,043	Sta = 'low'	NORMAL
0,110	0,026	Sta = 'low' and Oth = 'high'	NORMAL
0,130	0,293	Oth = 'normal'	HIGH
0,101	0,216	Sta = 'normal' and Oth = 'normal'	HIGH
0,084	0,000	Iss = 'high' and Sta = 'normal'	HIGH

Table 7.5: Induced rules with the Tertius algorithm, where the consequent is an influence class. I= confirmation value of the rule, II= the observed frequency of counter-instances of the rule in the data set.

Table 7.5 shows that the fraction TAS unit label distribution sums to one for all the individual influence type categories. This means that if one particular TAS unit class has a relatively low fraction, another class automatically has a relatively high fraction. From Table 7.5, one can distill that it seems that a high 'Issue' frequency in combination with a low 'Other' frequency seems to be more representative for highly influential people. People of low influence, on the other hand, score high on the 'Other' units and low on the 'Issues'. As could be expected from the confirmation values, post-hoc statistical analysis revealed that these hypotheses do not prove to be statistically significant (cf. Section 7.2.1).

A second experiment was performed with a data set containing the influence values added to all TAS unit labels and its associated features (including the relation that attaches the node to the tree). All of the features were again binned into the three (high, normal and low) bins. The top three rules[4] that were induced from the data for each influence category are reported in Table 7.6.

From the deduced rules depicted in Table 7.6 one could derive that relatively

[4](Note that confirmation rank is dependent on the number of features and that rankings between tables cannot be compared in a sense that one rule will be better than another.)

I	II	Antecedent	Consequent
0,079	0,009	ORT = 'low' and LL = STA and LL2 = YIS	HIGH
0,079	0,009	LL = STA and LL2 = YIS and NS = 'high'	HIGH
0,076	0,003	QMT = 'high' and L = 'low' and node = STA	HIGH
0,094	0,007	L = 'low' and rel = OPT and DB = 'normal'	NORMAL
0,091	0,007	QMT = 'low' and rel = OPT and DB = 'normal'	NORMAL
0,091	0,007	rel = OPT and DB = 'normal' and node = STA	NORMAL
0,093	0,621	QMT = 'low' and ORT = 'low' and L = 'low'	LOW
0,088	0,593	QMT = 'low' and ORT = 'low' and LPS* = 'low'	LOW
0,086	0,635	ORT = 'low' and L = 'low' and LPS* = 'low'	LOW

Table 7.6: The top three induced rules with the Tertius algorithm where the consequent is an influence class. I= confirmation value of the rule, II= the observed frequency of counter-instances of the rule in the data set. * LPS = length previous segment.

high influential people respond to people who provide responses to yes/no issues. People with a relatively low influence level seem to use fewer question marks, use the word 'or' less frequently and provide relatively short responses. This seems to align with the finding reported above that influential participants seem to raise more 'issues' and generally provide less units that can be labelled as 'other'.

7.5 Taking it all together

Given the results from the statistical investigations, the results on the classification performance and the rules that were inducted, one could try to construct a tentative profile of how influential participants, as experienced by actual meeting participants, distinguish themselves from less influential participants. When considering the previous sections, one could say that:

- Influential participants seem to raise more issues.
- Influential participants leave the provision of options, or possible solutions, to others.
- Influential participants seem to provide more evaluative information with respect to the contributions of others.
- Influential participants seem to respond to statements from others that follow after Yes/No Issues.
- Influential participants significantly elicit and provide more suggestions for action over the course of a meeting.

- Influential participants significantly provide less back-channels over the course of a meeting.
- Influential participants seem to provide less 'other' TAS units.
- Influential participants provide fewer unfinished utterances, or speech fragments over the course of a meeting.
- Influential participants seem to resort later in a discussion to shorter turns.

So it seems that if a participant raises issues, elicits solutions, evaluates these solutions and then steers towards a choice amongst the possible solutions, one indeed gets an intuitive sense of a person who is highly influential, and who controls the course of discussion. On the other hand, if someone just provides options, backchannels a lot to others, resorts to shorter contributions in the decision phase of the discussion indeed, then an intuitive profile of a less influential participant appears.

Exploitation of these profiles and the interrelation between both phenomena, however, do not prove to be sufficiently distinctive, in such a way that cross-fertilization of (features of) phenomena can yield machine learning algorithms to significantly improve their recognition performance. This result underlines that features have to correlate more than slightly with the phenomena of interest and also that 'just adding' features to the data set does not automatically improve the performance, in a sense that complementarity also plays a part.

Chapter 8

Future Meetings and Meeting Technology

"Assimilation into the Borg Collective might be inevitable, but we can still make it a more human place to live."
(Alex Pentland, 2005)

8.1 Introduction

Along with advanced meeting recognition technologies, such as described in the previous chapters, the innate human need for connectedness will drive science into the next decades. As humans will always be social animals that like to get together in groups, the challenge to realize the ability to be connected to anyone, anytime and anything is likely to be taken on.

The technological trend that started with the invention of the telegraph, or maybe even earlier with the invention of writing, will increasingly enable us to bring geographically dispersed people together by arranging a meeting of minds (see also Section 3.5.3). It is not unlikely that one day this meeting of minds, composed of humans, their representatives or completely autonomous systems, will take place in a meeting space that is able to match itself to the meeting type and exists both in reality as well as in virtuality. The increasing level of assistance from computers in the environment will make meetings more productive. The technology described in the previous chapters could result, for instance, in more balanced meetings and yield more creative ideas that will be worked out better. The meeting environment will in general understand much more of our daily activities and it will not be long before humans start to engage in interaction with the environment itself.

This chapter will shed some light on the pace and the implications of the possible technological developments for the meeting domain. Section 8.2 will discuss the current possibilities for achieving remote presence in both reality and in virtuality. Section 8.3 then elaborates on a special kind of meeting assis-

Figure 8.1: A futuristic 3D holographic representation of a meeting participant.

tant, the virtual chairman. The virtual chairman is an example application for the developed technologies throughout this thesis. This chairman should theoretically be able to aid meetings in any smart meeting space in real time. As sensors and associated smart systems are introduced at a greater scale, privacy, security, and other issues will undoubtedly start to play a part. This section also reports on a Wizard-of-Oz experiment that was conducted to test the expectation and the responses of meeting participants in relation to various sorts of meeting assistants, such as those described in this thesis. Section 8.4 then elaborates on a more general level about the ethical challenges and implications of the developed, and to be developed, meeting technologies. The chapter finishes with an overview of the most important technological challenges that lie ahead in the area of human computing in general, and meeting supporting systems in particular.

8.2 Remote presence

The availability of the other meeting participants is an important aspect before one can arrange a meeting among distributed people. Instant messaging systems, for example, generally have 'status' settings that reveal information about the others being available or not. Bentley et al. (2003) mentions amongst various types of awareness, the awareness of the availability of the other as one of the most important cues for starting a meeting. One can sense this way whether people are available to meet or not. Asynchronous awareness applications and devices have been around for almost two decades and vary in their realizations from lamps that change colors dependent on the mood of a family member (Tollmar and Persson, 2002), to flowers that bloom once a relative is at home[1]. For a meeting to take place, however, synchronous, rather than asynchronous awareness is indispensable. A meeting space does not come into being unless synchronous awareness is realized. Of course, one can listen in, but a real sense of remote presence has no chance of being achieved unless one can act on the

[1] http://web.media.mit.edu/ stefan/hc/projects/one2one/

environment such that a response can be measured.

Section 3.2 showed two models that can be used to categorize meetings. Meetings were divided in time and space (Figure 3.1) along the continuum between reality and virtuality (Figure 3.2). A model that combines these two, with the focus on synchronous communication, has been proposed by Benford et al. (1998) and is shown in Figure 8.2.

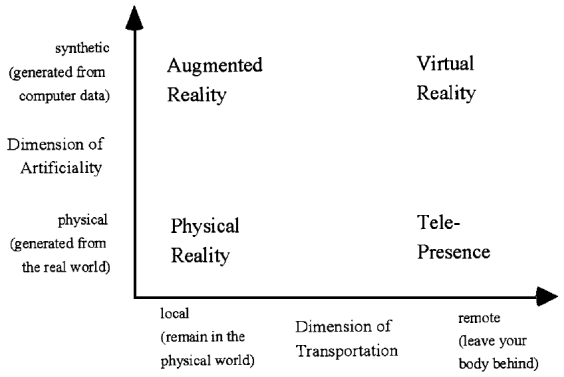


Figure 8.2: Classification of shared spaces according to Benford et al. (1998).

The two dimensions of the figure concern on the one hand the extent to which a group of participants (and objects) leave their local space behind and enter into some new remote space in order to meet with others (transportation) and on the other hand the extent to which the space is either artificial or is based on the physical world (artificiality). Transportation also includes the possibility of introducing remote participants and objects into the local environment as well as that it considers how groups of participants, and possibly other objects such as physical documents, might be transported and manipulated together (see e.g. Sakong and Nam (2006)). The two dimensions constitute four possible areas where meetings can take place: Reality, Augmented Reality, Virtual Reality and Tele-Presence. The remainder of this section will illustrate the three areas other than 'reality' by providing some examples of ongoing work.

Augmented reality

Shared augmented reality meetings supplement the users immediate physical surroundings with additional synthetic information, such as projections and annotations. An example of a future augmented reality meeting is shown in Figure 8.1. This figure shows an intergalactic meeting where people gather from all over the universe in order to discuss certain topics. The person in the middle is projected by imaginary 3D holographic technology. This 3D broadcasting allows the other participants to retain gaze and posture effects during communication and in essence makes it look as if the person really is *there*. Although this pic-

ture is taken from a fantasy movie, as computer and graphics hardware become more powerful and faster networks become available, new technologies from the domain of mixed or augmented reality can be used to add extra realism to video conferencing systems, and also to real life environments.

Nguyen et al. (2005) for instance show that when wearing semi-transparent glasses one can project captured humans in 3D at a rate of 25 frames per second into a mixed reality environment. A more or less similar system, but now tailored for conferencing purposes is depicted in Figure 8.3. This system, described in Billinghurst et al. (2002), allows any number of remote users to appear as life-sized images in a mixed reality environment. These virtual video windows can be placed anywhere around the user in space and spatial cues can this way be maintained. Theoretically the image planes could also move around their Y-axis (cf. Vertegaal (1998)) to allow support for natural gaze cues.

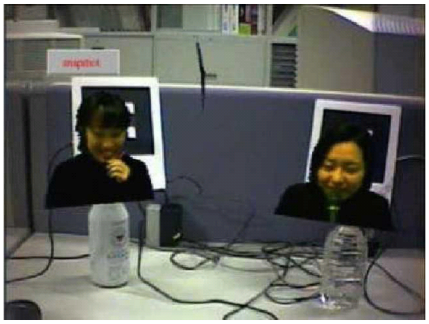


Figure 8.3: Augmented reality conferencing. Taken from Billinghurst et al. (2002).

Another trend, that can also be placed in the category of augmented reality, has to do with new volumetric three dimensional display devices, such as those described in (Favalora, 2005). These display devices make it possible to display participants in a meeting from different locations as if they were in the same room. Figure 8.4(a) shows an example of such a device: the teleorb[2]. The tele-orb falls in between the holographic projection and the system that overlays video images in the real world. The tele-orb is able to display 3D objects, such as faces and complete embodiments of humans and avatars, within an enclosed glass dome.

[2]taken from the classic adventure game 'The return to Zork'

(a) The TeleOrb (b) The Perspecta (c) Real Time 3D regeneration (d) 3D regeneneration

Figure 8.4: Towards the Teleorb

Recently Orbons (2006) demonstrated that the tele-orb is about to become reality. He showed that it is possible to create a 3D reconstruction of a human head at high resolution (256x256 pixels) at a rate of 16 frames per second by using multiple cameras at a time. An example of such a reconstruction is shown in Figure 8.4(c). These 3D models can in turn be transferred on high speed networks and displayed on 3D displays such as the Hitachi transpost[3] or the Perspecta display from Actuality Systems[4] (see Figure 8.4(b)).

Virtual Reality

Virtual Environments, such as the virtual meeting room that was described in Section 4.4.3, are composed of synthetic 3-D geometry possibly combined with texture-mapped images or video streams. When these environments visually replicate, or regenerate, the actions that are sensed at the sites of the remote participants a shared simulated environment emerges that makes it appear as if everyone is in the same room. Dependent on the sensed information one can regenerate the remote participant. In an approach described by Nijholt et al. (2005) only higher level information, such as body posture and gaze direction, is transmitted from the remote sites. The information is locally transposed onto embodiments that represent the participants. Another approach, that requires a little more bandwidth, reconstructs the view of the local participant from a visual hull that is computed remotely from a set of multi-view images (Laurentini, 1994). The result of this technique is shown in Figure 8.5.

An example of the possibilities offered by a virtual reality meeting environment is based on the fact that different meeting participants need not necessarily all have the same view of the virtual environment (see also Section 3.5.3). This could for instance mean that different participants can have a different perception of the seating arrangement so that they all feel more comfortable. For other

[3] http://hhil.hitachi.co.jp/products/transpost-e.html
[4] http://www.actuality-systems.com

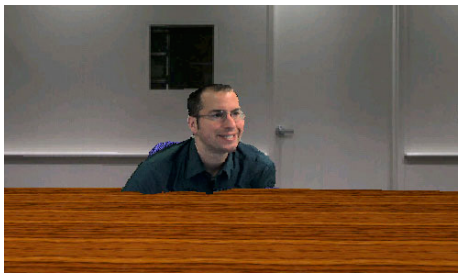


Figure 8.5: A real 3D representation in a virtual world. Taken from Slabaugh et al. (2002).

related issues, such as the influence of the submitted channels on the meeting itself see Section 3.5.

The most conspicuous aspect of these virtual reality systems is that, one day, it might become impossible for a human to distinguish whether the driving factor behind the virtual participants is a human, an interpretation of a human, or a completely autonomous system. Human participants will be aided by software systems that 'do' meetings for them, possibly by pretending to be the real human participant. (The actual appearance of any participant, be it a human, or an autonomous system, is in any virtual world obviously totally free to choose, see also Section 3.2.1 and 3.3.2.) This way the human can pursue its daily routine without being interrupted by a boring and lengthy meeting. If the assisting software runs short of knowledge how to deal with an unexpected situation, it will always have the possibility to sign in its principal, give a brief update of the situation, and hand over the control of the representation to the remotely residing human.

Tele-presence

Tele-presence refers to a user interacting with another real, rather than virtual, place. Tele-presence and presence in virtual reality both came into being through the transmission of the (representation of) sensed environments at either side of the connection. The main functional difference is the entity on the other end: a real environment in the case of tele-presence, versus a virtual environment virtual reality. For tele-presence, it is generally a robot that acts on the sensed movements at the remote site. This robot can be equipped with the ability to manipulate objects and even to conduct surgeries. Haptic, or tactile force, feedback is given to the user, so that he or she can get some approximation of the weight, stiffness, size, and/or texture of the remote objects.

Two examples of tele-presence robots that can be used in a meeting context are the Giraffe (Figure 8.6(b)) and the Pebbles Robot (Figure 8.6(a)). In a sense both robots act as a stand-in for the user. People near the robot can see and hear the user and interact with him or her as if the user were truly present.

(a) Pebbles (b) The Giraffe

Figure 8.6: Tele-presence robots

Pebbles was claimed by its developer, Telbotics, to be the world's first fully functioning 'tele-presence' application (Fels et al., 1999). Pebbles was created to connect hospitalized, homebound and special needs children to their home classroom, allowing for participation in classroom activities and social contact. The robot was first used in 1997 at Toronto's Hospital for Sick Children. The Giraffe Video Conferencing Robot, developed by HeadThere[5] is a more or less similar mobile robot that can be moved around its location by remote control using the internet. Both Pebbles and the Giraffe allow a user to hear, see, and speak at a remote site.

8.3 The virtual chairman

Autonomous software systems are, along with the emergence of advanced recognition technology for human interaction, such as shown in this thesis, likely to increasingly influence the course of our daily meetings. The introduction of embodied conversational agents that stand in for real meeting participants is already taking place (Nijholt et al., 2005). One step away, and also not unlikely, is that autonomous systems will take over certain participant roles, such as the role of the meeting chairman.

In a meeting, a chairman has to manage the meeting process in order to maximize the output of the meeting, stick to the agenda and to maintain a positive meeting atmosphere. An autonomous system that replicates the chairman should preferably carry out all of these activities. Obviously, some of these are far more easy to realize than others. Guarding agenda and time constraints is,

[5] www.headthere.com

for instance, a far more straightforward task than taking care of the decision-making process, not to mention trying to exploit the expertise of all present meeting participants to realize maximal synergy.

Based on an analysis of what is going on in the meeting, a virtual chairman could influence and steer the progress of the meeting (request a vote, encourage silent people to speak, mention gaps in the argumentation, etc.). The better the state of the recognition technology, the more enhanced the tasks of the chairman can hypothetically be. If this technology becomes advanced enough, a chairman that is able to detect potentially tense situations could, for example, try to defuse such situations by making a joke, or changing the subject of the discussion.

Figure 8.7: An impression of a virtual meeting chaired by a virtual chairman.

The automatic detection of a dominance hierarchy could enable a virtual meeting chair to maintain a more balanced meeting and the availability of argumentation structures could inform a chairman about unexplored ideas and topics of contention. To actually realize this type of support radical new ways of analyzing discourse content are necessary. As Purver et al. (2005) argue, this problem is much harder than conventional discourse analysis in human-machine dialogues, since the computer cannot steer the process by posing the right questions, or try to understand something by initiating a clarification dialogue.

Pertaub et al. (2002) showed that people can be influenced in their behavior as well as their assessment of a situation by the presence of a virtual audience composed of autonomous agents that respond to their behavior. In work from DiMicco (2004) a system called Second Messenger is described that shows real-time text summaries of participants contributions. After increasing the visibility of the less frequently speaking group members, it appeared that these started to speak more frequently than before, whereas the more dominant people started to speak 15% less. Both examples show that it indeed seems possible to build systems that are capable of influencing the meeting process. To my knowledge, hardly any other experiments have been conducted yet to see how and whether the presentation of retrieved meeting information can actually impact a meet-

ing. The remainder of this section describes a summary of experiments that investigated whether, and in what form, meeting assistants such as a meeting chairman, that aim at improving meeting effectiveness, can work in practice[6].

8.3.1 Putting live assistance to the test

Central in the experiments that were conducted was the question of how and for which aspects an assistant, such as a meeting chairman, should act in order to be listened to whilst in the mean time not being too intrusive. In the experiments, the meeting assistants were simulated using a Wizard of Oz technique. This means that the meeting participants were led to believe that they interacted with an autonomous system, when in fact a human being controlled the behavior of the system remotely. This approach was chosen because an implementation of a complete assistant in the first place would be technically too time consuming, if not (as yet) impossible and secondly it was expected that a good Wizard-of-Oz experiment would yield nearly identical results.

The research setting The research setting that was used is displayed in Figure 8.8(b). The experimenter monitored the meeting room and controlled the assistant in the remote control center. Live video footage of the meeting was displayed on the screen in the control center and a ceiling mounted microphone was used to capture the audio. The actions of the 'assistant' were realized via a monitor in front of the meeting chair and/or a speaker set that resided inside the meeting room. The consistency of the experiment was guaranteed by the creation of a script that the experimenter followed.

The experiment As a preliminary investigation questionnaires were issued, to 15 different people who were known to regularly seat a meeting, in order gain some insights into the meeting aspects that were considered useful for a hypothetical meeting assistant; 9 were fully completed and returned. The most notable aspects that were mentioned were off-topic, balance and time information. The chairmen also expected that information presented on a display would be more beneficial to the meeting efficiency in comparison to voiced information. The screen was also expected to be less intrusive than the voice-over (see Figure 8.8(a)). With this information different systems with varying intrusiveness levels were composed for the experiment. Table 8.1 shows descriptions of some of these systems ranked from least to most intrusive according to the perceptions expressed in the questionnaires. Two student committees (of eight and seven members respectively) were subsequently exposed to all versions of the system over a period of four weeks. Before each meeting participants were asked to provide the agenda, an expected time-line, the names of participants and the seats they would occupy during the meeting. After each meeting questionnaires were issued in order to discover how the assistant and its actions were received by the meeting participants. Participants were asked, amongst other things, to

[6] See Kuperus (2006) and Broenink (2006) for a more elaborate description.

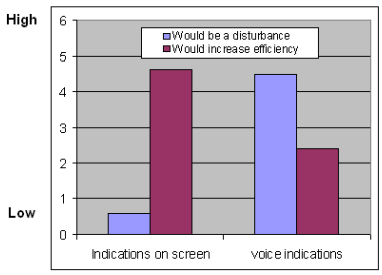
rate from high to low, on a seven point scale, their perception of the meetings' efficiency, the meeting being off-topic, the meeting being balanced as well as their perception of the system being enjoyable, disturbing and intrusive. After every session the chairmen were also again asked to give their opinion about the disturbance and efficiency for both the voice as well as the screen feedback strategies.

System	Description
1	Displays messages on a screen when an item is due to be finished in five, two or zero minutes. Also displays messages when something is off-topic, a subject takes too long or when a discussion is unbalanced.
2	Similar to system 1, but instead of displaying messages, continuously displaying a clock.
3	Similar to system 1, but instead of displaying messages, voice samples were played.

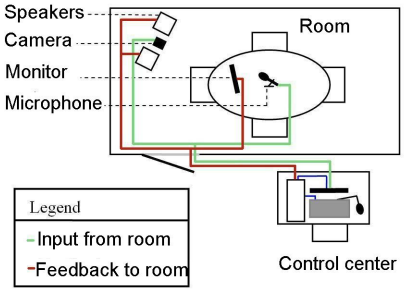
Table 8.1: Description of the systems simulated for the experiment .

Some findings and results When considering the participants' ratings of degree of intrusiveness versus efficiency, Figure 8.8(d)) shows that the added intrusiveness of System 3 pays off in terms of meeting efficiency. Notable is the fact that the perceived meeting disturbance for System 2, does not seem to be higher than for System 1, as expected by the chairmen beforehand. So active messages do not seem to be more intrusive than an omnipresent static clock. System 3 shows a much higher efficiency increase along with the increased intrusiveness and a slight increase in meeting disturbance. The enjoyability for System 3 however is rated much lower than for Systems 1 and 2 (see Figure 8.8(c)). It also appeared that, in contrast to the pre-meeting questionnaire results, the chairmen now rated the Systems 2 and 3 equal with respect to the perceived efficiency. Voice messages were still found more intrusive than the text messages.

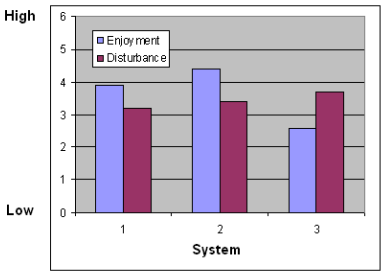
An interesting side result was that when the system used voiced feedback, the participants of the meeting appeared to be much more aware of their own behavior. When they tended to go off-topic, the participants corrected themselves very quickly, sometimes saying: "off-topic", before continuing with the current agenda item. It should be noted that, although the above findings speak in favor of a system that assists the meeting process, much additional research is required, for instance by examining a larger number of groups over a longer period of time.



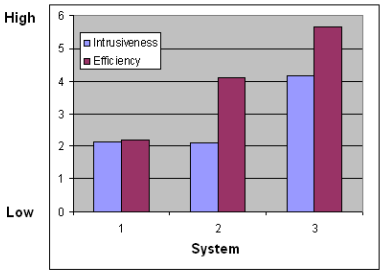
(a) Some of the Chairmen Questionnaire results.



(b) Experiment Setup



(c) Enjoyment vs. Disturbance as rated by the participants



(d) Intrusiveness vs. Efficiency as rated by the participants

Figure 8.8: Some results of the Wizard of Oz experiment

8.4 Ethical implications and considerations

Ongoing developments in the area of progressing meeting technology in general and human computing in particular could result in far reaching ramifications for human life and human well-being. The advent of the networked society, has permitted people to interact with each other remotely in a fashion unprecedented in history. This, on the one hand, has brought about enormous benefits and convenience, whilst on the other hand, it has extended a dark side where a new technology is abused or disrupts human relations (Nishida, 2007).

Socially aware communications systems quickly bring to mind the country of Oceania where everybody is under complete surveillance by the authorities[7]. Not to mention the fact that more and more signals emitted by humans are increasingly being stored and are available for post-hoc analysis. A serious and very important consideration is how far one wants to go with developing these sorts of technologies. I want to be absolutely clear that with the research described in this thesis I do not have, and hopefully never will have, any intentions in any of these frightening directions at all. I sincerely hope that future (meet-

[7]See Orwell (1984).

ing) technologies will only be used to the benefit of all of human kind and our surrounding environment. I also strongly agree with Pentland (2005) who stated that "systems that are aware of human social signaling, and that adapt themselves to human social context are expected to leave us with organizations that are not only more efficient, but that also better balance our formal, informal, and personal lives."

It is however not unlikely that the introduction of new technology in the meeting domain will, for example, pose difficult challenges for participants and their supervisors. Although a participant's access to remote participants all over the globe, for instance, may theoretically increase his or her productivity, ubiquitous connections to others comes along with temptations for distraction and the wasting of time. Not to mention the temptation that will emerge for supervisors to implement automated supervision techniques. How useful would it be for an employer to gain automatic insights into the performance of his or her participants over the previous meetings? And what would the participants think of this? It seems not unimaginable that these 'monitoring' techniques could lead to tension, distrust, and resentment. So what could seem beneficial and an advantage at first sight, might turn out to be a disadvantage in the end.

Another potential danger that lies enclosed in emergent technologies is overreliance on systems that are not flawless and that are trained on a specific domain. Overreliance on automatic systems, especially without knowledge of the rationale behind the systems could lead to annoying situations in which high expectations can turn out to become nasty dampers. The impact for meeting technology will perhaps not be as large as for an earthquake warning system that makes a mistake, but for a business meeting where large interests are at stake I assume it would be better to at least think twice and to always refrain from blindly following a system's proposals, and rather consider its advice as suggestions that could be taken into account. Of course the level of authority and autonomy that is given to the system plays a part in this. Also, as the technologies have been trained for a specific domain, the risk exists that they are put into practice in different domains. The models in this thesis have for example been trained on four person meetings and as a consequence they cannot be blindly transferred to a six person meeting, nor can one predict how these systems will deal with unseen situations.

The last issue I want to raise here is what would happen when more and more meetings become virtual meetings, rather than face-to-face meetings in reality. In the past people and companies might have picked their partners based on their evaluation of charisma, reliability, personal and communicative skills, and apparent competence. Today, however, the look and feel of a web site, chat environments, several emails and perhaps a phone call take the place of the search for personal and professional compatibility. What are the consequences of this, apart from the obvious reasons that one extends its working range? How reliable is the party at the other end, how sincere are his or her intentions, and is he or she not pretending to be someone else (cf. Bailenson et al. (2004))?

8.5 Challenges ahead

The characteristics of emerging HCI systems for the meeting domain imply new approaches to usability engineering and associated evaluation and testing techniques. Emerging meeting systems that are devised to support, and to a certain degree also understand, meetings as they naturally occur, require the ability to comprehend messages emitted through various signals, including voice, gestures, gaze and facial expressions. When allowing humans to communicate naturally to the input devices of human computing systems, these systems should be able to distill within this gamut of signals all the items that are to the system's interest. Despite considerable research effort in the field of multi-modal fusion (see e.g. Oviatt (2003)), knowledge about how humans combine different channels is still limited. Not to mention the recognition of the behavior of the group as a whole. Furthermore, the system should also be sufficiently prominent, because a lack of a prominence might result in users who are unaware of the system's existence (see e.g. Nijholt et al. (2004)).

The first questions that one could probably ask oneself is: What sort of (human computing) systems should one build for the meeting domain? and Why? In my case I have tried to devise two models that show insights into the meeting process. Insights, on top of which systems can be built that should aid the meeting process, such that users of the technology obtain a competitive advantage. However, as we have seen in Chapter 2, the meeting domain is quite a broad domain and indeed, what can be beneficial for a particular aspect can result in negative consequences for other aspects (i.e. a gain in meeting efficiency decreases the meeting's enjoyableness). So the answer to this question is not trivial. From a market perspective one could interview potential customers for their demands. Here one comes across the problem that users usually have no idea what to demand for, as they are unaware of the technological possibilities and generally satisfied with the current system that they are used to. To be honest, from my point of view the technological pull for the development of meeting supporting technologies seems more and more to turn into a technological push. Especially since the greatest time and money saver, the connection of geographically dispersed people by means of voice, has long been realized. It was not until mobile broadband services appeared that subsequent steps could be set.

The aim of this section is to shed some more light on the main challenges that lie ahead for new and to be developed systems in the meeting domain. The remainder of this section will address both the perception side, the training time and the output generation side.

8.5.1 Appropriate input perception

Meeting data that are obtained from sensors, such as microphones and cameras, need to be sensed by sufficiently accurate sensors. The subsequent recognition module that transforms the perceived data into information should, in turn, also be sufficiently reliable for its task. As we have seen in the previous chapters

current state-of-the art, in basically every automatic observation and classification task that concerns human behavior, including speech, is far from errorless recognition. Error-less recognition is a serious challenge that exists in many human computing domains, especially since in the human case, subtleties easily go unnoticed, the environment may be noisy, unseen situations might emerge and the currently applied technology is error prone. One way to assess the input reliability is by using benchmark sets. Well-known benchmark sets are, for example, the NIST RT sets (Fiscus et al., 2006) for automatic speech recognition or FRVT and FRGC for face recognition (Phillips et al., 2006). These sets are specific for a given context and task. Since they contain ground truth and the error metrics are known, they allow for good comparison. However, they still evaluate only the reliability of the input to the system, rather than the evaluation of a system as a whole. One way to recover from errors is through the usa of repair mechanisms. A system could for example rephrase what was understood before it takes action. Feedback or insight in the system's state could be useful in this sense. Still, there are many challenges that are to be resolved before these sorts of repair mechanisms will be in general use (cf. Bellotti et al. (2002)).

It is often mentioned that human behavior is to be interpreted in a given context. For example, a smile in an everyday conversation can be a sign of appreciation, whereas, during negotiation, it can be a sign of disagreement. So, for the reliable interpretation of human behavior, it is important for human sensing systems to be aware of the context of the situation. To date, there is no consensus on what context precisely is, or on how we should specify this (van Bunningen et al., 2005). Without a good representation for context, developers are left to develop *ad hoc* systems for storing and manipulating this key information (see e.g. Abowd and Mynatt (2000)). Usually, the context is specified as the identity and location of the users in combination with the characteristics and timing of the action performed. Ideally, however, the history and the intentions of the user are also to be taken into account. Sometimes the major components of context are referred to as the 5 W's (Abowd and Mynatt, 2000): who, what, where, when, why. It is difficult to automatically assess the values for most, if not all, of these properties. Intille et al. (2003) observe for smart homes that the user naturally considers contexts that a system will not be able to obtain, they therefore propose to use supportive systems as suggestive, rather than pro-active.

8.5.2 Task composition and evaluation

For the meeting domain, many factors can be identified that can have a substantial impact on the success of applications that are not easily quantified and therefore hard to evaluate. A challenge that exists for the tasks that are to be performed by meeting assisting applications, is in the first place to devise this set of tasks. The tasks that one could envision for meeting supporting technologies quickly result in unrealistic, and often too abstract dreams. One could, for example, envision tasks in the line of meeting atmosphere regulation,

the enhancement of the users meeting experience and the establishment of enduring social relationships. However, none of these dreams can be easily and objectively achieved, neither via explicit tasks, nor in terms of right and wrong. Whittaker et al. (2000), in addition to this, observed that many developed HCI systems can be considered radical inventions that just do not build further on established knowledge about user activities, tasks and techniques, but rather push the technology envelope and invent new paradigms. The absence of a strictly defined and easily measurable task makes it hard to determine when a new solution is better, rather than different (Newman, 1997).

It is therefore a real challenge to come up with a set of tasks for a system that on the one hand is really useful for both the participant and the meeting and on the other hand can be implemented given the current state of technology. Microsoft Powerpoint is, next to (video) conferencing applications and all sorts of amplifying auxiliaries such as microphone-speaker combinations, light bulbs and heating devices, perhaps the best example of real beneficial and working meeting technology. It not only allows users to communicate much more information through a variety of modalities to the other meeting participants, but it is also actually working robustly and everywhere.

Another aspect that has not been mentioned is that developed systems should be evaluated in a context that is as close as possible to the context of authentic use (Abowd and Mynatt, 2000). The evaluation of meeting supportive systems in controlled laboratory settings is, for example, likely to cause unnatural behavior of the participants and the users (cf. Section 1.3.1). Another drawback of using laboratory testing is that parameters can be controlled (background noise, lightning conditions) that cannot be controlled in the context of authentic use, resulting in an insuperable situation. For more information about evaluation aspects of HCI applications in general refer to Poppe et al. (2007).

8.5.3 Appropriate output generation

Also the output that a system should provide in terms of messages that it sends to the user is an issue that is far from being solved. The process of choosing and combining modalities to best convey the intended message is central for multimodal output generation. This is a complex and highly knowledge-intensive process that depends on the type of information that has to be conveyed, the intention the application has with the information, the specifics of the context and of course the user (Bachvarova et al., 2007). Proper understanding and modelling of the nature of each modality is a task that is increasingly gaining attention (see Bateman et al. (2001); Wainfan and Davis (2004)).

The area of embodied conversational agents, for example, is a typical domain that is confronted with a severe lack of proper evaluation (Xiao et al., 2002). Many people believe that such interfaces have great potential to be beneficial in HCI. However, as conversational agents, for example, might indeed seem to be the most 'natural' interface (as the user does not have to learn complex command structures and functionality), users, on the other hand now also start to expect other human functionalities to be implemented in these system. The

problem is that this embodied agent thus should not only be able to interpret an almost infinite lexicon, including aspects such as intonation, gaze patterns, facial expressions and gestures, but in addition, is now also expected to conduct this behavior itself. I think that a huge challenge lies in finding the appropriate means of conveying information from a system to a user. Ideally, a system should be able to select the modality that suits the user the most. Be it either a diagram, a voice sample, or a complete 3D animation. It should not matter. As in human-human interaction, the sender should adapt to the user and whilst sending the message, continuously sense what part of the emitted message has already been received and understood.

On the generation side in order to convey a message in a meeting context, a system should address the right people whilst employing the appropriate systematics associated with the desired behavior that has been selected in order to convey the message. But what if an embodied conversational agent intends to be 'influential' in a meeting, in a sense similar to the notion of influence described in Chapter 4? In this context it is important to note that these systematics cannot be mapped one-to-one onto the features that were selected on the recognition side. The research described in Chapter 4 and 5 of this thesis just concerned the detection side of two concepts. And although I have tried to bring together as many features as possible to eventually end up with the most predictive features that can be used by a system in order to, to a certain extent, assess these concepts, it is from my opinion very unlikely that these feature sets will ever achieve a notion of the associated 'concept' (for which these features work at the detection side) at a users-end. Not in the first place, because some of the features use historical and contextual information, but mainly because the features that work well on the detection side, might just not work on the generation side. One reason why it might not work is the context of use of the system. What might work for one user, might just not work for another user and the same holds for the environment.

Chapter 9

Conclusions

9.1 Findings

Meetings are an extremely complex phenomenon where many aspects of everyday life play a part and come together. Meetings are often inefficient and expensive. Factors of meeting success relate mostly to the extent to which the meeting expectations that exist beforehand are achieved. Meeting preparation, control and the willingness to contribute are central. Meetings should be balanced and not be monopolized by one or two dominant people as this inhibits the participation of others and the creativity to generate ideas or solutions. Technology has improved meetings ever since the invention of telegraphy in the 1850's. It was shown that the ability to interact remotely is thus far perhaps one of the greatest technological achievements for the meeting process. Several applications have been shown that are beneficial to meetings, before, during and after their occurrence.

Systems that are able to perceive and understand what is going on in a meeting pertain to the emergent human computing paradigm in which adaptive systems respond in accordance to their perceived (human)environment. The methodology of corpus based research investigates the possibilities for this technological trend to sense higher level concepts after a clever combination of more direct observations. This methodology requires a model that describes the phenomena that should be recognized as well as a carefully chosen example domain on which this model should be manually applied. After manual application machine learning algorithms can be trained in order to replicate the human observations from a set of features that are both easily observable and expected to relate the phenomena under consideration.

This methodology has been applied for both the phenomenon of argumentation structures, as they unfold in meetings, and influence hierarchies as they remain in the minds of the meeting participants. Influence hierarchies that classified meeting participants from four-person meetings into the classes of high, medium and low influential appeared to be replicable in around 70% of the cases

tested. The most important features in order for a participant to be recognized as influential in this case appeared to be that the participant had to take a lot of long turns, he or she had to initialize a lot of topics, or be the meeting chair. For argumentation structures a model (TAS) was defined that represents the argumentation during discussions in a graphical manner by means of a tree graph that preserves the discussion flow. The extent in which two classification tasks, out of the five main tasks that need to be executed before complete automatic recognition is established, could be successfully performed was examined. It appeared that the TAS-node labels could be correctly reproduced for around 75% of the cases and that predefined relations amongst these nodes could be labelled correctly for around 60% of the cases. For both the phenomena of influence hierarchies and argument structures, applications have been constructed that aim to assist post-hoc meeting analysis.

After combining the data that was used to detect the influence hierarchies and argumentation structures, that via rule induction and statistical analysis some cross-links between both phenomena could be identified, though not all statistically significant. From these interdependencies a profile of an influential participant who is engaged in a discussion could be constructed. According to the data, the profile of an influential meeting participant in a discussion comprises a participant who raises a lot of issues, elicits solutions, evaluates these solutions and then steers towards a choice amongst these solutions. The interdependencies between both phenomena, however, did not prove useful for the mutual classification tasks. The findings show that a slight, though significant, difference in feature value distributions alone is not in every case a sufficient prerequisite for a feature to be useful in a classification task.

The future of meeting technology will encompass both synthetic and physical meetings that can take place locally as well as remotely. Augmented reality, telepresence and virtual reality meeting supporting techniques have been discussed. The VMR, a virtual replica of the AMI meeting room in Martigny was created (see Figure 4.2). Apart from being used as a virtual meeting place, its possibilities for meeting augmentation with all sorts of deduced phenomena, such as addressee, gaze and also influence information, have been shown. The VMR can further be used to test human observation capabilities for annotation schema creation and function as a test environment for autonomous software agents. The impact that autonomous software systems might have on the meeting process has been tentatively investigated by means of a wizard of Wizard-of-Oz experiment. It appeared that an increase in meeting effectiveness came along with a decrease in the meeting's enjoyableness.

9.2 Implications and interpretations

Turning back to the question what people really want from meetings, and how meetings can be improved by technological means and more specifically the technology developed in this thesis; one could ask oneself if the expected break through in everyday meeting conduct can indeed be realized by the current state-

of-the-art in technology, as was hypothesized at the beginning of this thesis. More specifically, in our case, we could ask ourselves what the benefits are, that a system equipped with current state-of-the-art technology might bring as a result of the recognized phenomena of argument structure and influence hierarchy in order to improve meetings.

Whenever both phenomena become available, applications could potentially do a lot of good things to a meeting. In the case of influence hierarchies, for example, the knowledge could be exploited by future systems that strive for a more balanced meeting process. In the hope that the system's actions result in behavioral changes, the meeting chair, the influential, or the non-influential participants could collectively, or individually, be informed about the system's findings. But also once the meeting is over, the detected hierarchies contain valuable information about participant behavior that can be used for training and optimization strategies. The availability of the argument structures, on the other hand, reveal the 'trains of thought' that were followed, and point out issues that have not been sufficiently addressed. On top of this structure other algorithms could be built that analyze the meeting on a more semantic level. To avoid redundancies, these algorithms could, for example, try to relate issues that are currently being addressed to issues that were addressed in previous meetings. Algorithms might autonomously crawl for background information in available resources, such as the web, to inform participants about controversies, or chances to win the debate. Meetings will improve, as the quality of the discussions increases and the participants do not need to be afraid of being shouted down when providing ideas, or solutions. This, combined with the fact that post-hoc meeting browsers can preserve the knowledge that has been debated and make this information easily available to the interested public, the benefits seem eminent.

The truth, however, shows that with the given approach the systems were able to detect the phenomena which we were after only to a *reasonable* extent. The question now arises if, in the first place, one can ever achieve these benefits at all, and in the second place, to what extent the achieved recognition levels of our algorithms can contribute to these hypothesized goals. Thereby also taking into account that the Wizard-of-Oz experiment showed that the increase on the time aspect, came at the cost of the meeting's enjoyableness.

When trying to formulate an answer, one should realize that in the first place we clearly have just set some explorative steps in the human computing domain. A domain where the technological developments are far from error free and that by itself is still in its infancy. I am of the opinion that the results described in this thesis are encouraging in a sense that they can be used in suggestive systems that, for example, can suggest label categories for annotators or behavioral changes for meeting participants. But I also have tried to make clear that blind reliance on this technology might lead to erroneous decision making and that the temptation of abuse can lead to nasty privacy and responsibility issues. So at this moment this technology can, hinging on its performance, in the best case be used as suggestive, or informative. This by itself is not a bad achievement, especially when we realize that decisions concerning higher level human-human

communication phenomena, such as those that occur in meetings, are of a highly subjective nature on which humans themselves often disagree.

9.3 Scientific contributions

Apart from the fact that the AMI corpus that was used is the first large scale multi-modal meeting corpus on which statistical research in the meeting domain has become possible, and that the domain of this thesis, meetings, is widely known, there undoubtedly exists a widespread encouragement for any meeting supporting type of technology. The following scientific contributions I find worth mentioning:

- A vast amount of literature has been reviewed in a way that a comprehensive introduction into the subject of meetings, meeting influence and argumentation has been realized.
- An annotation schema and associated annotation tools to structure argumentation in a discussion have been developed and are available to the public[1].
- TAS annotations are available for the whole AMI hub corpus.
- The implementation of two browser-plugins for the JFerret Meeting browser has been realized.
- The findings described add substantially to the understanding of how more and less influential participants can be automatically detected in a smart meeting environment
- The findings described add substantially to the understanding how argument structures can be automatically obtained.

9.4 Limitations

A number of important limitations need to be considered. First and foremost is perhaps the fact that all our results are based on four-person remote control design meetings that, although they are as natural as one can get, all follow a predefined script. This has the potential consequence that participant behavior, as exposed in these meetings, might very well differ in other meetings. It was, for example, already mentioned in Section 2.4 that the composition and the size of the group have a direct impact on the behavior of the participants. But also the fact that people had to follow a predefined script that somehow constrained their behavior could have played a part, just as the fact that the participants were informed that they were 'recorded'. As a result of this, caution must be applied, as findings might not be transferable to meetings of a different flavor.

[1] See http://wwwhome.cs.utwente.nl/~rienks

Second is the fact that the experiments, and most of the feature collection, were started from provided transcriptions. Given this fact, combined with the performance drops on ASR output as described in Section 6.5.3, means that one cannot conclude that the results described in this thesis can directly be transferred to real-time applications. Section 5.6.1, on the other hand, showed the implementation of a model that, with a careful selection of predictive and direct obtainable features from the audio signal, a useful real-time application can be constructed. The performance of this 'derived' model approximates the performance that can be obtained when using all features.

As third and final point, I need to mention the sample size that was used for these experiments. Although I am unaware of any comparable experiments of this scale, the total number of samples from a machine learning point of view is relatively small and a larger data set is always likely to yield more reliable results. Associated with this fact is the reliability from the manual annotations. We did the best we could to assess some sort of reliability measure as well as an objective annotation scheme. The Virtual Kappa measure, as introduced in Section 6.3.4, showed some encouraging results, but due to the sparsity of time and money it just was impossible to have all annotations double checked, or performed by more than one annotator. And of course, the more the annotators agree, the more objective the observation gets. But inter-annotator disagreement can, apart from inter-annotator subjectivity, also be attributed to the (low) quality of the annotation schema. It therefore always remains a trade-off where to start when willing to improve observations. Machine learning algorithms will, however, always be searching for commonalities and correlations and the more the annotators disagree, the lower the recognition rate of the algorithms will be.

9.5 Reconsiderations and Future work

Looking back at the experiments conducted, one can always ask oneself the question: what should one do differently if a chance appeared to re-do the experiments?

In the case of the obtainment of the class labels for both phenomena it would, as mentioned, have been better to use more annotators, and to have worked with more annotated data so that we could have obtained more reliable results. Then, one could have used different features, maybe not so much based on the transcriptions, but rather those features that can be obtained directly from the raw data sources. This way, the achieved accuracy would also better reflect the performance that a system would achieve in real-time operation. Whereas in the current situation, one could say that we leaped beyond the current state of technology by using the transcriptions as input source. So a focus on real-time applications, rather than post hoc analysis could be a direction to think of.

Also, more experiments could have been conducted with respect to examining the eventual impact of the recognition of the phenomena on the meeting process. And even though the experiments reported in Section 8.3.1 showed an increase in meeting effectiveness, larger scale testing is required to see whether

these tentative conclusions can be further grounded. One could, for example, also experiment with various ways of selectively displaying the information to just the more influential, or the less influential participants in order to investigate if and where such a system can achieve its greatest impact.

Another point concerns the data set that was investigated in order to see whether and how the phenomena of influence hierarchy and argumentation structure are related. Looking back, it would have been wiser to have all the meetings annotated with influence information for which the TAS annotations existed or vice versa. But as this experiment was not planned at the time that the annotations were created we missed the boat, in a sense that our data set potentially could have been larger. A related subject that has not been investigated is how influential participants influence the content of the debate. We found that influential people steer towards a solution, but how is this done? Do they resort to other than rational argumentation techniques in order to persuade the other participants, in a sense that they abuse their position, or is it on other grounds? In this respect, a deeper semantic analysis of the argumentation process might yield interesting insights.

Stepping back, and considering the challenges that exist on the path towards the full recognition of higher level communication phenomena as sketched in Section 8.5, the most important question that remains to be answered, from my point of view, is the question if the foreseen technological benefits in the domain of human computing are not just false hopes, in a sense that the dream just cannot not be realized. Especially since human communicative behavior is so complex that one can truly ask the question if systems will ever be able to understand (parts of) this behavior autonomously and reliably at all. The thing is merely that, similar to emotion recognition, the recognition task for human communication could prove too complex and too subjective, in a sense that there exists no right or wrong, and that, as a result, unless asking for confirmation, a system can never be sure if it correctly understands what is going on. This, in turn, does certainly not align with the initial thoughts on human computing. At this moment in time, and although I have contributed just a very small piece of the greater picture, I do think, that the results shown in this thesis suggest that it is not impossible to one day have systems that analyze the influence levels of the meeting participants and show the argument structures of debates. Especially since research in the human computing area is still in its infancy, we possibly have just scratched the surface of future human computing possibilities. I am therefore surely looking forward, but also with a little irony, to the day that the Perkomat will hit the market and becomes available universally.

.

Summary

Meetings are often inefficient. They are numerous and unavoidable. If we look at the technological developments in this area we quickly see that along with the introduction of the microphone and the data projector, the execution of a meeting for the participants has become much more easy. Yet there are still many aspects of a meeting that can be improved, where technology in its current stage has not contributed much. There is for instance hardly any technology that is able to autonomously interpret, or analyze, aspects of the meeting process.

An automatic analysis of a meeting could provide valuable insights for both the attendants, as well as for those interested parties who could not attend. These insights hypothetically could in turn lead again to more successful meeting processes. It is, for example, often the case that one or two dominant participants can monopolize a complete meeting and thereby prohibit others from contributing. Another example is that the argumentation that has been put forward and that led to a certain decision is often forgotten and lost, not to mention that during a discussion just one line of argumentation can be in the center of attention.

It is investigated to what extent the latest technological developments can provide automatic insights into both, so-called higher-level meeting phenomena. To enable the automatic recognition, a descriptive and computationally accessible model has been created for the phenomenon of dominance hierarchy as well as for argument structure. Although the model for a dominance hierarchy did not require more than a ranking of the participants, the model that describes the argumentation structure requires interpretation of the individual contributions, as well as the knowledge of how to label contributions in the context of the discussion.

From a social psychological background, correlated and more easily detectable aspects and signals that either have proven to be, or were expected to be useful for the recognition of the phenomena have been collected. The resulting corpus of meeting recordings in combination with the collected relevant aspects was, combined with human interpretations of the phenomena presented, used as input for machine learning algorithms. These algorithms were trained on this data with the aim to have them learn how to replicate the human interpretations of these higher level phenomena on unseen data.

For a dominance hierarchy this appears to be possible in approximately 70% of the meetings that were tested. For the recognition of an argumentation

structure, the individual contributions can be correctly interpreted in terms of the predefined model in around 75% of the cases. The contextual labels that describe the relations between these contributions can be correctly replicated in around 60% of the cases. All in all this shows that full reproduction of human interpretation is not (yet) completely possible, but although the results are not perfect, the recognition is already sufficient for the creation of at least some meeting supporting applications. The future will show if and how these technological possibilities can eventually lead to the enhancement of meeting processes and if the meeting experience for all participants can, as intended, be improved.

.

Bibliography

Abowd, G. D. and Mynatt, E. D. (2000). Charting past, present, and future research in ubiquitous computing. *ACM Transactions on Computer-Human Interaction*, 7(1):29–58.

Agrawal, R., Imielinski, T., and Swami, A. (1993). Mining association rules between sets of items in large databases. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 207–216.

Al-Hames, M. Dielmann, A., Gatica-Perez, D., Reiter, S., and Renals, S. Zhang, D. (2005). Multimodal Integration for Meeting Group Action Segmentation and Recognition. In *Proceedings of the second conference on Machine Learning and Multi-Modal Interactions (MLMI)*.

Al-Hames, M. and Rigoll, G. (2005). A multi-modal mixed-state dynamic bayesian network for robust meeting event recognition from disturbed data. In *Proceedings 6th International Conference on Multimedia and Expo, ICME 2005*, Amsterdam, The Netherlands.

Alben, L. (1996). Quality of experience: defining criteria for effective interaction design. *Interactions*, 3(3):11–15.

Alvehus, J. (1999). Meeting metaphors, a study of narratives about meetings, theis no. 753. Master's thesis, Department of Computer and Information Science, Linnkoping University.

Ang, J., Liu, Y., and Shriberg, E. (2005). Automatic dialog act segmentation and classification in multiparty meetings. In *Proceedings of the 30th International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Antunes, P. and Carrio, L. (2003). Modeling the information structures of meetingware. In *Proc. of Workshop de Sistemas de Informao Multimdia e Cooperativos (COOP-MEDIA'03)*.

Aoki, P., Romaine, M., Szymanski, M., Thornton, J., Wilson, D., and Woodruff, A. (2003). The mad hatters cocktail party: A social mobile audio space supporting multiple simultaneous conversations. In *Proc. ACM SIGCHI Conf. on Human Factors in Computing Systems*, pages 425–432. ACM Press.

Appelbaum, E et Batt, R. (1994). *The New American Workplace : Transforming the Work System in the United States*. Cornell University/IRL Press, Ithaca, New York.

Argyle, M. and Dean, J. (1965). Eye-contact, distance and affiliation. *Sociometry*, 28:289–304.

Argyle, M., Ingham, R., Alkema, F., and McCallin, M. (1973). The different functions of gaze. *Semiotica*, 7:19–32.

Austin, J. (1962). *How to do things with words*. Oxford University Press.

Bachvarova, Y., van Dijk, B., and Nijholt, A. (2007). Towards a unified knowledge-based approach to modality choice. In *Proceedings Workshop on Multimodal Output Generation (MOG 2007)*, pages 5–15.

Bailenson, J., Beall, A., Loomis, J., Blascovich, J., and Turk, M. (2004). Transformed social interaction: Decoupling representation from behavior and form in collaborative virtual environments. *Presence*, Vol. 13, No. 4, August:428441.

Bailenson, J., Blascovic, J., Beall, A., and Loomis, J. (2001). Equilibrium theory revisited: Mutual gaze and personal space in virtual environments. *Presence: Teleoperators and Virtual Environments*, 10(6):583–598.

Bakeman, R. and Gottman, J. (1997). *Observing interaction: An introduction to sequential analysis, Second Edition*. Cambridge University Press, Cambridge, UK.

Baldridge, J. and Lascarides, A. (2005). Annotating discourse structures for robust semantic interpretation. In *Proceedings of the Sixth International Workshop on Computational Semantics (IWCS)*, Tilburg, The Netherlands.

Bales, R. (1950). *Interaction Process Analysis*. Addison-Wesley.

Bales, R. and Cohen, S. (1979). *SYMLOG: A System for the Multiple Level Observation of Groups*. The Free Press.

Bales, R., Strodbeck, F., Mills, T., and Roseborough, M. (1951). Channels of communication in small groups. *American Sociological Review*, 16:461–468.

Banerjee, S. and Pedersen, T. (2003). The design, implementation, and use of the Ngram Statistic Package. In *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, Mexico City.

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 Tenth International Conference on Human-Computer Interaction*.

Barakonyi, I., Frieb, W., and Schmalstieg, D. (2003). Augmented reality videoconferencing for collaborative work. In *Proceedings of the 2nd Hungarian Conference on Computer Graphics and Geometery*.

Barthelmess, P. and Ellis, C. A. (2005). The neem platform: An evolvable framework for perceptual collaborative applications. *Journal of Intelligent Information Systems*, 25(2):207–240.

Barzilay, R. (1997). Lexical chains for summarization. Master's thesis, Ben Gurion University of the Negev, Department of Mathematics and Computer Science.

Basu, S., Choudhury, T., Clarkson, B., and Pentland, A. (2001). Learning human interaction with the influence model. Technical Note 539, MIT Media Laboratory.

Bateman, J. A., Kamps, T., Kleinz, J., and Reichenberger, K. (2001). Constructive text, diagram and layout generation for information presentation: the dartbio system. *Computational Linguistics*, Vol. 27(4):409449.

Batliner, A., Steidl, S., Schuller, B., Seppi, D., Laskowski, K., Vogt, T., Devillers, L., Vidrascu, L., Amir, N., Kessous, L., and Aharonson, V. (2006). Combining efforts for improving automatic classification of emotional user states. In *Proceedings of the 5th Slovenian and 1st International Language Technologies Conference (IS-LTC 2006)*.

Baumeister, R. F. and Leary, M. R. (1995). The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin*, 117:497–529.

Bellotti, V., Back, M., Edwards, K., Grinter, R. E., Jr., A. H., and Lopes, C. V. (2002). Making sense of sensing systems: Five questions for designers and researchers. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI'02)*, pages 415–422, Minneapolis, MN.

Benford, S., Greenhalg, C., Reynard, G., Briwn, C., and Koleva, B. (1998). Understanding and constructing shared spaces with mixed reality boundaries. *ACM Transactions on Computer-Human Interaction (ToCHI)*, 5(3):185–223.

Benford, S., Schndelbach, H., Koleva, B., Anastasi, R., Greenhalgh, C., Rodden, T., Green, J., Ghali, A., Pridmore, T., Gaver, B., Boucher, A., Walker, B., Pennington, S., Schmidt, A., Gellersen, H.-W., and Steed, A. (2005). Expected, sensed, and desired: A framework for designing sensing-based interaction. *ACM Transactions on Computer-Human Interaction*, 12(1):3–30.

Bentley, F., Tollmar, K., Demirdjian, D., Koile, K., and Darrell, T. (2003). Perceptive presence. *IEEE Computer Graphics and Applications*, 23(5):26–36.

Berger, J., Cohen, B., and Zelditch, M. (1966). *Siciological Theories in Progress, Volume I*, chapter Status Characteristics and Expectation States, pages 29–46. Houghton Mifflin, Boston, USA.

Berger, J., Ridgeway, C., and Zelditch, M. (2002). Construction of status and referential structures. *Sociological Theory*, 20(2):157–179.

Berger, J., Rosenholtz, S., and Zelditch Jr, M. (1980). Status organizing processes. *Annual Review of Sociology*, 6:479–508.

Berry, P., Gervasio, M., Uribe, T., Pollack, M., and Moffitt, M. (2005). A personalized time management assistant: Research directions. In Saphiro, D., editor, *Persistant Assistants: Living and working with AI, workshop at the AAAI Spring Symposium 2005*. AAAI Press.

Billinghurst, M., Cheok, A., Prince, S., and Kato, H. (2002). Real world teleconferencing. *IEEE Computer Graphics and Applications*, 22(6):11–13.

Bilmes, J. (2000). Dynamic bayesian multinets. In *Proceedings of the 16th Conference in Uncertainty in Artificial Intelligence*, Stanford University, CA, U.S.A.

Bluedorn, A. C., Turban, D., and Love, M. (1999). The effects of stand-up and sit-down meeting formats on meeting outcomes. *Journal of Applied Psychology*, 84(2):277–285.

Blumer, A., Ehrenfeucht, A., Haussler, D., and Warmuth, M. (1987). Occam's razor. *Information Processing Letters*, 24(6):377–380.

Boser, B., Guyon, I., and Vapnik, V. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the 5th Annual ACM Workshop on COLT*, pages 144–152. ACM Press.

Brants, T., Skut, W., and Uszkoreit, H. (2003). *Syntactic annotation of a German newspaper corpus*, chapter 5. Kluwer, Dordrecht (NL).

Briggs, R., de Vreede, G., and Nunamaker Jr., J. (2003). Collaboration engineering with thinklets to pursue sustained success with group support systems. *Journal of Management Information Systems*, 19(4):31–64.

Briggs, R. and Vreede, G. (2001). Thinklets: Achieving predictable, repeatable, patterns of group interaction with group support systems (gss). In *Proceedings of the 34th Hawaii International Conference on System Sciences*.

Broenink, E. (2006). How a computer actor influences the time-efficiency of a meeting. In *Proceedings of the 4th Twente Student Conference on IT*. Twente University Press.

Bruggen, J. v. (2003). *The use of external representations of argumentation in collaborative problem solving*. PhD thesis, Open Universiteit Nederland.

Buckingham Shum, S. (1997). Negotiating the construction and reconstruction of organisational memories. *Journal of Universal Computer Science*, 3(8):899–??

Buckingham Shum, S. (2003). *Visualizing Argumentation: Software Tools for Collaborative and Educational Sense-Making*, chapter The Roots of Computer Supported Argument Visualization. Springer Verlag, London, UK.

Bunt, H. (1979). Conversational principles in question-answer dialogues. *The theory of questions*.

Burleson, C. (1990). *Effective Meetings: The Complete Guide*. John Wiley & Sons.

Burroughs, W., Schultz, W., and Autrey, S. (1973). Quality of argument, leadership votes, and eye-contact in three person leaderless groups. *Journal of Social Psychology*, 90:89–93.

Burstein, J., Marcu, D., and Knight, K. (2003). Finding the write stuff: Automatic identification of discourse structure in student essays. *IEEE Intelligent Systems*, 18(1):32–38.

Butt, D. and Fiske, D. (1968). Comparison of strategies in developing a scale for dominance. *Psychological Bullitin*, 6:505–519.

Carberry, S. (1989). A pragmatics-based approach to ellipsis resolution. *Computational Linguistics*, 15(2):75–96.

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

Carletta, J. (2006). Unleashing the killer corpus: experiences in creating the multi-everything ami meeting corpus. In *Proceedings of the Language Resources and Corpora (LREC)*.

Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kraaij, W., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska, A., McCowan, I., Post, W., Reidsma, D., and Wellner, P. (2005). The ami meetings corpus. In *Proceedings of the Measuring Behavior 2005 symposium on "Annotating and measuring Meeting Behavior"*. AMI-108.

Carletta, J., Isard, A., Isard, S., Kowtko, J., Doherty, G., and Anderson, A. (1997). The reliability of a dialogue structure coding scheme. *Computational Linguistics*, 23(1):13–32.

Carlson, L., Marcu, D., and Okurowski, M. (2001). Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*, pages 1–10, Morristown, NJ, USA. Association for Computational Linguistics.

Carver, C. and Scheier, M. (1990). Origins and functions of positive and negative affect: A control-process view. *Psychological review*, 97:19–35.

Chapanis, A., Ochsman, R., Parrish, R., and Weeks, G. (1972). Studies in interactive communication. : The effects of four communication modes on the behavior of teams during cooperative problem-solving. *Human Factors*, 14:487–509.

Chen, H., Finin, T., and Joshi, A. (2004). A context broker for building smart meeting rooms. In *Proceedings of the Knowledge Representation and Ontology for Autonomous Systems Symposium, (AAAI Spring Symposium)*.

Chen, H. and Perich, F. (2004). Intelligent agents meet semantic web in a smart meeting room. In *Proceedings of the Third International Joint Conference on Autonomous Agents & Multi Agent Systems (AAMAS 2004)*.

Clancey, W. (1997). *Situated Cognition: On Human Knowledge and Computer Representations*. Cambridge University Press.

Clark, H. (1996). *Using language*. Cambridge University Press.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 1(20):37–46.

Cohen, S. (1993). *Organizing for the future: the new logic for managing complex organisations*, chapter New approaches to teams and teamwork, pages 194–226. Jossey-Bass Publishers, San Francisco.

Conklin, J. and Begeman, M. (1988). gibis: a hypertext tool for exploratory policy discussion. *ACM Trans. Inf. Syst.*, 6(4):303–331.

Cook, P., Ellis, C., Graf, M., Rein, G., and Smith, T. (1987). Project nick: meetings augmentation and analysis. *ACM Trans. Inf. Syst.*, 5(2):132–146.

Cooley, J. (1959). On mr. toulmin's revolution in logic. *Journal of Philosophy*, 56:297–319.

Coon, C. (1946). The universality of natural groupings in human societies. *Journal of Educational Sociology*, 20:163–168.

Core, M. G. and Allen, J. F. (1997). Coding dialogues with the DAMSL annotation scheme. In Traum, D., editor, *Working Notes: AAAI Fall Symposium on Communicative Action in Humans and Machines*, pages 28–35, Menlo Park, California. American Association for Artificial Intelligence.

Corston-Oliver, S. (1998a). *Computing Representations of the Structure of Written Discourse*. PhD thesis, University of California, Santa Barbara.

Corston-Oliver, S. (1998b). Identifying the linguistic correlates of rhetorical relations. In Stede, M., Wanner, L., and Hovy, E., editors, *Proceedings of the COLING-ACL*, pages 8–14, Somerset, New Jersey. Association for Computational Linguistics (ACL).

Coser, R. (1955). Laughter among collegues. *Psychiatry*, 23:81–95.

Cristea, D., Ide, N., Marcu, D., and Tablan, V. (1999). Discourse structure and co-reference: An empirical study. In Cristea, D., Ide, N., and Marcu, D., editors, *Proceedings of the Workshop on the Relationship Between Discourse/Dialogue Structure and Reference, Association for Computational Linguistics ACL'99*, pages 46–53. New Brunswick, New Jersey.

Dahlgren, K. (1988). *Naive Semantics for Natural Language Understanding*. Boston: Kluwer Academic Press.

Danninger, M., Flaherty, G., Bernardin, K., Ekenel, H., Khler, T., Malkin, R., Stiefelhagen, R., and Waibel, A. (2005). The connector - facilitating context-aware communication. In *Proceedings of the International Conference on Mumtimodal Interfaces*, Trento, Italy.

Davies, N. and Gellersens, H.-W. (2002). Beyond prototypes: Challenges in deploying ubiquitous systems. *IEEE Pervasive Computing*, 2(1):26–35.

De Vreede, G., Vogel, D., Kolfschoten, G., and Wien, J. (2003). Fifteen years of GSS in the field: A comparison across time and national boundaries. In *Proceedings of the 36th Hawaii International Conference on System Sciences*. IEEE Press.

Dennis, A., George, J., Jessup, L., Nunamaker Jr, J., and Vogel, D. (1988). Information technology to support electronic meetings. *MIS Quarterly*, 12(4):591–624.

Dielmann, A. and Renals, S. (2004). Dynamic bayesian networks for meeting structuring. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

DiMicco, J. (2004). Designing interfaces that influence group processes. In *Doctoral Consortium Proceedings of the Conference on Human Factors in Computer Systems (CHI 2004)*.

DiMicco, J., Pandolfo, A., and Bender, W. (2004). Influencing group participation with a shared display. In *CSCW '04: Proceedings of the 2004 ACM conference on Computer supported cooperative work*, pages 614–623. ACM Press.

Dovidio, J. and Ellyson, S. (1985). *Power, Dominance, and Nonverbal behavior*, chapter Patterns of Visual Dominance Behavior in Humans, pages 129–149. Springer Verlag.

Doyle, M. and Straus, D. (1976). *How to make Meetings Work*. Berkely Publishing group.

Drew, J. (1994). *Mastering Meetings: Discovering the Hidden Potential of Effective Business Meetings (second edition)*. McGraw Hill-companies.

Duda, R., Hart, P., and Stork, D. (2000). *Pattern Classification, Second Edition*. John Wiley & Sons.

Dunbar, N. and Burgoon, J. (2005). Perceptions of power and interactional dominance in interpersonal relationships. *Journal of Social and Personal Relationships*, 22(2):207–233.

Duncan, S. (1972). Some signals and rules for taking speaking turns in conversations. *Journal of Prsonality and Social Psychology*, 23(2):283–292.

Duncan, S. and Niederehe, G. (1974). On signalling that it's your turn to speak. *Journal of Experimental Social Psychology*, 10:234–247.

Edwards, J. A. (2001). *Handbook of Discourse*, chapter Transcription in Discourse, pages 321–348. Mass: Blackwell Publishers.

Ehrlich, S. (1987). Social and psychological factors influencing the design of office communications systems. In *Proceedings of the SIGCHI/GI conference on Human factors in computing systems and graphics interface*, pages 323–329.

Ellis, C. and Barthelmess, P. (2003). The Neem dream. In *Proceedings of the 2003 conference on Diversity in computing*, pages 23–29. ACM Press.

Ellis, C., Barthelmess, P., Quan, B., and Wainer, J. (2001). Neem: An agent-based meeting augmentation system. Technical report, University of Colorado at Boulder,Department of Computer Science. CU-CS-937-02.

Ellis, C., Wainer, J., and Barthelmess, P. (2003). *Agent supported collaborative work*, volume 8 of *Multi Agent Systems, Artificial Societies and Simulated Organizations*, chapter Agent Augmented Meetings. Springer Verlag.

Engelbart, D. (1963a). *Computer Supported Cooperative Work: A Book of Readings*, chapter A conceptual Framework for the Augmentation oof Man's Intellect, Reprint in. Morgan Kaufman 1988, San Mateo, CA.

Engelbart, D. (1963b). *Vistas in information handling*, chapter A conceptual framework for the augmentation of man's intellect, pages 1–29. Spartan Books, Washington D.C., U.S.A.

Erol, B., Lee, D., and Hull, J. (2003). Multimodal summarization of meeting recordings. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME 2003)*.

Favalora, G. (2005). Volumetric 3d displays and application. *Computer*, 38(8):37–44.

Fels, D., Williams, L., Smith, G., Treviranus, J., and Eagleson, R. (1999). Developing a video-mediated communication system for hospitalized children. *Telemedicine Journal*, 5(2):193–207.

Fernandez, R. and Picard, R. (2002). Dialog act classification from prosodic features using support vector machines. In *Proceedings of speech prosody 2002*.

Fiscus, J. G., Radde, N., Garofolo, J. S., Le, A., Ajot, J., and Laprun, C. (2006). The rich transcription 2005 spring meeting recognition evaluation. In Renals, S. and Bengio, S., editors, *Revised Selected Paper of the Machine Learning for Multimodal Interaction Workshop 2005 (MLMI'05)*, volume 3869 of *Lecture Notes in Computer Science*, pages 369–389, Edinburgh, UK.

Fisek, M. and Ofsche, R. (1970). The process of status evolution. *Sociometry*, 33:327–346.

Fisher, S. S., McGreevy, M., Humphries, J., and Robinett, W. (1986). Virtual environment display system. In *In Symposium on Interactive 3D Graphics*, pages 77–87.

Flach, P. and Lachiche, N. (2002). Confirmation-guided discovery of first-order rules with tertius. *Machine Learning*, 1(42):61–95.

Galaczy, P. (1999). Electronic meeting systems: Win-win group decision making? Technical report, Queen's University.

Galley, M., McKeown, K., Fosler-Lussier, E., and Jing, H. (2003). Discourse segmentation of multi-party conversation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics. (ACL2003)*.

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*, pages 669–676.

Garofolo, J. S., Laprun, C. D., Michel, M., Stanford, V. M., and Tabassi, E. (2004). The nist meeting room pilot corpus. In *Proc. of the LREC2004*.

Garrido, L. and Sycara, K. (1995). Multi-agent meeting scheduling: Preliminary experimental results. In Lesser, V., editor, *Proceedings of the First International Conference on Multi-Agent Systems (ICMAS'95)*. The MIT Press: Cambridge, MA, USA.

Gatica-Perez, D. (2006). Aalyzing group interactions in conversations: A review. In *Proceedings of the IEEE Conference on Multisensor Fusion and Integration*.

Gatica-Perez, D., McCowa, I., Zang, D., and Bengio, S. (2005). Detecting group-interest level in meetings. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICCASP)*, Philadelphia.

Gaver, B. and Martin, H. (2000). Alternatives: exploring information appliances through conceptual design proposals. In *Proceedings of the conference on Human factors in computing systems (CHI'00)*, pages 209–216.

Girgensohn, A., Boreczky, J., and Wilcox, L. (2001). Keyframe-based user interfaces for digital video. *Computer*, 34(9):61–67.

Godfrey, J., Holliman, E., and McDaniel, J. (1992). Switchboard: Telephone speech corpus for research and development. In *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 517–520, San Francisco.

Goetsch, G. and McFarland, D. (1980). Models of the distribution of acts in small discussion groups. *Social Psychology Quarterly*, 43(2):173–183.

Goffman, E. (1955). On face-work: An analysis of ritual elements in social interaction. *Psychiatry: Journal of Interpersonal Relations*, 18(3):213–231.

Goffman, E. (1974). *Frame analysis: An essay on the organization of experience.* Harper and Row.

Gordon, M. (1985). *How to Plan and Conduct a Successful Meeting.* Sterling Publishing Co.

Gratch, J., Rickel, J., Andre, E., Badler, N., Cassell, J., and Patajan, E. (2002). Creating interactive virtual humans: Some assembly required. *IEEE Intelligent Systems*, 17(4):54–63.

Greenhalgh, C. and Benford, S. (1995). Virtual reality tele-conferencing: Implementation and experience. In *Proc. Fourth European Conference on Computer Supported Cooperative Work (ECSCW95).*

Grice, H. (1975). *Logic and conversation*, chapter Syntax and Semantics: Speech Acts, pages 41–58. Academic Press.

Grosz, B. and Sidner, C. (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Grudin, J. (1988). Why CSCW applications fail: problems in the design and the evaluation of organizational interfaces. In *Proceedings of the Conference on Computer Supported Cooperative Work (CSCW'88)*, pages 85–93, New York, USA.

Grudin, J. (1994). Groupware and social dynamics: Eight challenges for developers. *Communications of the ACM*, 37(1):93–105.

Habermas, J. (1984). *The Theory of Communicative Action, Volume 1: Reason and the Rationalization of Society.* Beacon Press.

Hall, M. (1999). *Correlation-based Feature Subset Selection for Machine Learning.* PhD thesis, Department of Computer Science, University of Waikato, Waikato, N.Z.

Halliday, M. and Hassan, R. (1976). *Cohesion in English.* Longman.

Halliday, M. and Hassan, R. (1985). *An Introduction to Functional Grammar.* Edward Arnold Press.

Hassine, A., Defago, X., and Ho, T. (2004). Agent-based approach to dynamic meeting scheduling problems. In *Proceedings of the 3rd International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS'04).* IEEE society.

Hearst, M. (1997). Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64.

Hellriegel, D., Slocum Jr., J., and Woodman, R. (1995). *Organizational Behavior, seventh edition.* West publising company.

Henley, N. (1972). *Body Politics: Power, sex and nonverbal Communication.* Prentice-Hall.

Hersey, P. and Blanchard, K. (1988). *Management of Organizational Behavior: Utilizing Human Resources.* Prentice Hall.

Hewitt, C. (1977). Viewing control structures as patterns of passing messages. *Artificial Intelligence*, 8(3):323–364.

Heylen, D. (2006). Head gestures, gaze and the principles of conversational structure. *International Journal of Humanoid Robotics*, 3(3):241–267.

Hiltz, S. and Turoff, M. (1978). *The network nation: Human communication via computer.* Addison-Wesley Publishing Company, Inc., Massachusetts.

Hirokawa, R. and Pace, R. (1983). A descriptive investigation of the possible communication based reasons for effective and ineffective group decision making. *Communication Monographs*, 50(4):363–379.

Hirschberg, J. and Litman, D. (1994). Empirical studies on the disambiguation of cue phrases. *Computational Linguistics*, 19(3):501–530.

Hobbs, J. (1979). Coherence and coreference. *Cognitive Science*, 3(1):67–90.

Hocking, D. (1996). *Hockings Rules: The essential guide to conduction meetings*. Simon and Schuster.

Hoffmann, L. (1979). Applying experimental research on group problem solving to organizations. *Journal of applied behavioral science*, 15:375–391.

Hofman, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42:177–196.

Hoyle, R., Pinkley, R., and Insko, C. (1989). Perceptions of social behavior: Evidence of differing expectations for interpersonaland intergroup interaction. *Personality and Social Psychology Bulletin*, 15:365–376.

Hsueh, P., Moore, J., and Renals, S. (2006). Automatic segmentation of multiparty dialogue. In *Proccedings of the Eleventh conference of the European Association for Computational Linguistics EACL06)*.

Hsueh, S. and Moore, J. (2007). What decisions have you made?: Automatic decision detection in meeting conversations. In *Proceedings of the 1st meeting of the North America Chapter of the Association for Computational Liguists NAACL'07 Vol 1*. submitted.

Huang, J. and Zweig, G. (2002). Maximum entropy model for punctuation annotation from speech. In *Proceedings of the International Conference on Spoken Language Processing ICSLP*, page 917920, Denver, Colorado.

Ide, N. (2004). *Preparation and Analysis of Linguistic Corpora*. Blackwell Publishing).

Intille, S. S., Tapia, E. M., Rondoni, J., Beaudin, J., Kukla, C., Agarwal, S., Bao, L., and Larson, K. (2003). Tools for studying behavior and technology in natural settings. In *Proceedings of the International Conference on Ubiquitous Computing (Ubicomp'03)*, volume 3869 of *Lecture Notes in Computer Science*, pages 157–174, Seattle, WA.

Jaimes, A., Omura, K., Nagamine, T., and Hirata, K. (2004). Memory cues for meeting video retrieval. In *CARPE'04: Proceedings of the the 1st ACM workshop on Continuous archival and retrieval of personal experiences*, pages 74–85. ACM Press.

Jaimes, A., Sebe, N., and Gatica-Perez, D. (2006). Human centered computing: A multimedia perspective. In *Proceedings of the ACM MultiMedia'06*.

Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., and Wooters, C. (2003). The icsi meeting corpus. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 364–367.

Jebara, T., Ivanov, Y., Rahimi, A., and Pentland, A. (2000). Tracking conversational context for machine mediation of human discourse. In *In AAAI Fall 2000 Symposium - Socially Intelligent Agents - The Human in the Loop*.

Johansen, R. (1988). *Groupware: Computer Support for Business Teams*. The Free Press.

John, G. and Langley, P. (1995). Estimating continuous distributions in bayesian classifiers. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 338–345. Morgan Kaufman.

Jonassen, D. (2000). Toward a design theory of problem solving. *Educational Technology Research & Development*, 48(4):63–85.

Jones, D. (2006). I avatar: Constructions of self and place in second life. *Gnovis, Georgetown's peer reviewed Journal of Communication, Culture and Technology*.

Jones, D., Wolf, F., Gibson, E., Williams, E., Fedorenko, E., Reynolds, D., and Zissman, M. (2003). Measuring the readability of automatic speech-to-text transcripts. In *Proceedings of the European Conference on Speech Communication and Technology*, page 15851588, Geneva, Switserland.

Jovanovic, N., Op den Akker, R., and Nijholt, A. (2005). A corpus for studying addressing behavior in multi-party dialogues. In *Proc. of The sixth SigDial conference on Discourse and Dialogue*.

Jovanovic, N., Op den Akker, R., and Nijholt, A. (2006). Addressee identification in face-to-face meetings. In *11th Conference of the European Chapter of the ACL (EACL)*, Trento, Italy.

Ju, W. and Leifer, L. (To-Appear). The design of implicit interactions. *Design Issues, Special Issue on Design Research in Interaction Design*.

Jurafsky, D. and Martin, J. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall PTR, Upper Saddle River, NJ, USA.

Jurafsky, D. and Shriberg, E. Biasca, D. (1997). Switchboard swbd-damsl shallow-discourse-function annotation (coders manual,draft 13). Technical Report 97-02, University of Colorado, Institute of Cognitive Science.

Jurafsky, D., Shriberg, E., Fox, B., and Curl, T. (1998). Lexical, prosodic, and syntactic cues for dialog acts. In Stede, M., Wanner, L., and Hovy, E., editors, *Discourse Relations and Discourse Markers: Proceedings of the Conference*, pages 114–120. Association for Computational Linguistics, Somerset, New Jersey.

Kahn, J., Ostendorf, M., and Chelba, C. (2004). Parsing conversational speech using enhanced segmentation. In *Proceedings of Human Language Technology Conference / North American Chapter of the Association for Computational Linguistics annual meeting HLT-NAACL*, Boston, USA.

Kaiser, E., Demirdjian, D., Gruenstein, A., Li, X., Niekrasz, J., Wesson, M., and Kumar, S. (2004). Demo: A multimodal learning interface for sketch, speak and point creation of a schedule chart. In *Proceedings of the International Conference on Multimodal Interfaces - ICMI'04*, pages 329–330, State College, Pennsylvania, USA.

Kanselaar, G., Erkens, G., Andriessen, J., Prangsma, M., Veerman, A., and Jaspers, J. (2003). *Visualizing Argumentation: Software Tools for Collaborative and Educational Sense-Making*, chapter Desiging Argumentation Tools for Collaborative Learning. Springer Verlag, London, UK.

Kendon, A. (1967). Some functions of gaze-direction in social interaction. *Acta Psychologica*, 26:22–63.

Kestler, J. (1982). *Questioning Techniques and Tactics*. McGraw-Hill.

Kiesler, S., Siegel, J., and McGuire, T. (1984). Social psychological aspects of computermediated communication. *American Psychologist*, 39(10):1123–1134.

Kim, J. Woodland, P. (2001). The use of prosody in a combined system for punctuation generation and speech recognition. In *Proceedings of the European Conference on Speech Communication and Technology EUROSPEECH*, page 27572760, Aalborg, Denmark.

Kravitz, D. and Martin, B. (1986). Ringelmann rediscovered: The original article. *Journal of Personality and Social Psychology*, 50:936–941.

Kucera, H. and Francis, W. (1967). *Computational Analysis of Present-Day American English*. Brown University Press.

Kunz, W. and Rittel, H. (1970a). Issues as elements of information systems. Working Paper WP-131, Univ. Stuttgart, Inst. Fuer Grundlagen der Planung.

Kunz, W. and Rittel, H. W. J. (1970b). Issues as elements of information systems. Working Paper WP-131, Univ. Stuttgart, Inst. Fuer Grundlagen der Planung.

Kuperus, J. (2006). The effect of agents on meetings. In *Proceedings of the 4th Twente Student Conference on IT*. Twente University Press.

Kurohashi, S. and Nagao, M. (1994). Automatic detection of discourse structure by checking surface information in sentences. In *Proceedings of the 15th conference on Computational linguistics*, pages 1123–1127.

Lascarides, A., Asher, N., and Oberlander, J. (1992). Inferring discourse relations in context. In Thompson, H. S., editor, *Proceedings of the Thirtieth Annual Meeting of the Association for Computational Linguistics*, pages 1–8, San Francisco. Morgan Kaufmann.

Latan, B., Williams, K., and Harkins, S. (1979). Many hands make light the work: The causes and consequences of social loafing. *Journal of Personality and Social Psychology*, 37:822–832.

Lathoud, G., McCowan, I. A., and Odobez, J.-M. (2004). Unsupervised Location-Based Segmentation of Multi-Party Speech. In *Proceedings of the 2004 ICASSP-NIST Meeting Recognition Workshop*, Montreal, Canada.

Laurentini, A. (1994). The visual hull concept for silhouette-based image understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(2):150–162.

Lee, M. and Ofsche, R. (1981). The impact of behavioral style and status characteristics on social influence: A test of two competing theories. *Social Psychology Quarterly*, 44(2):73–82.

Leffler, A., Gillespie, D., and Conaty, C. (1982). The effects of status differentiation on nonverbal behavior. *Social Psychology Quarterly*, 45(3):151–161.

Leventhal, N. (1995). Using groupware to enhance team decision making. *Information Strategy*, 12(fall):6–13.

Licklider, J., Taylor, R., and Herbert, E. (1968). The computer as a communication device. *International Science and Technology*, 76:21–31.

Lisowska, A. (2003). Multimodal interface design for the multimodal meeting domain: Preliminary indications from a query analysis study. Technical report, ISSCO/TIM/ETI, Universit de Genve, Switserland. IM2.MDM Report 11.

Liu, Y. (2004). *Structural Event Detection For Rich Transcription Of Speech*. PhD thesis, Purdue University.

Lombard, M. and Ditton, T. (1997). At the heart of it all: The concept of presence. *J. Computer-Mediated Communication*, 3(2).

Loomis, J., Blascovich, J., and Beall, A. (1999). Immersive virtual environment technology as a basic research tool in psychology. *Behavior Research Methods, Instruments and Computers*, 31(4):557–564.

Mann, W. and Thompson, S. (1987). Rhetorical structure theory: A theory of text organization. Technical Report ISI/RS-87-190, University of Southern California.

Mann, W. and Thompson, S. (1988a). Rhetorical structure theory: toward a functional theory of text organization. *Text*, 8:243–281.

Mann, W. C. and Thompson, S. A. (1986). *Natural Language Generation: New Results in Artificial Intelliligence, Psychology and Linguistics*, chapter Rhetorical Structure Theory: Description and Construction of Text Structures, pages 279–300. Kluwer, Boston, USA.

Mann, W. C. and Thompson, S. A. (1988b). Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

March, J. and Sevon, G. (1984). Gossip, information and decision making. *Advances in Information Processing and Organizations*, 1:95–107.

Marcu, D. (1997a). The rhetorical parsing of natural language texts. In *Proceedings of the Meeting of the Association for Computational Linguistics (ACL)*, pages 96–103.

Marcu, D. (1997b). *The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts*. PhD thesis, Department of Computer Science, University of Toronto.

Marcu, D., Amorrortu, E., and Romera, M. (1999). Experiments in constructing a corpus of discourse trees. In *Proceedings of the ACL'99 Workshop on Standards and Tools for Discourse Tagging*, pages 48–57, Maryland.

Marcu, D., Carlson, L., and Watanabe, M. (2000). The automatic translation of discourse structures. In *Proceedings of the 1st meeting of the North America Chapter of the Association for Computational Liguists NAACL'00 Vol 1.*, pages 9–17.

Marcu, D. and Echihabi, A. (2002). An unsupervised approach to recognizing discourse relations. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*, Philadelphia, PA.

Mark, W., Randolph, S., Finch, M., Van Verth, J., and Taylor, R. (1996). Adding force feedback to graphics systems: Issues and solutions. *Computer Graphics*, 30(Annual Conference Series):447–452.

Martin, J. (1992). *English Text: System and Structure*. John Benjamins Publishing Co.

McCowan, I., Gatica-Perez, D., Bengio, S., Lathoud, G., Barnard, M., and Zhang, D. (2005). Automatic analysis of multimodal group actions in meetings. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(3):305–317.

McGrath, J. (1984). *Groups, Interactions and Performance*. Prentice Hall.

MCI WorldCom, 1998 (1998). Meetings in the UK : A study of trends, costs, and attitudes toward business travel and teleconferencing, and their impact on productivity. MCI WorldCom Conferencing Whitepaper.

McLuhan, M. (1964). *Understanding Media: The Extensions of Man*. McGraw Hill.

Meyer, S. and Rakotonirainy, A. (2003). A survey of research on context-aware homes. In *Proceedings of the Australasian information security workshop conference on ACSW frontiers 2003*, pages 159–168. Australian Computer Society, Inc.

Mignault, A. and Chaudhuri, A. (2003). The many faces of a neutral face: Head tilt and perception of dominance and emotion. *Journal of Nonverbal Behavior*, 27(2):111–131.

Milgram, P. and Kishino, F. (1994). A taxonomy of mixed reality visual displays. *IEICE Transactions on Information Systems*, E77-D(12).

Miller, G. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Mitkov, R. (2002). *Anaphora Resolution*. Longman.

Mongue, P., McSween, C., and Weyer, J. (1989). *A profile of meetings in corporate America: results of the 3M meeting effectiveness study*. Annenberg School of Communications, University of Southern California.

Monk, A. F., McCarthy, J., Watts, L., and Daly-Jones, O. (1996). *CSCW requirements and evaluation*, chapter Measures of process, pages 125–139. Springer-Verlag, Berlin, Germany.

Moore, J., Kronenthal, M., and Ashby, S. (2005). Guidelines for AMI speech transcriptions. Technical Report 1.2, IDIAP, Univ. of Edinburgh.

Moore, J. and Pollack, M. (1992). A problem for RST: The need for multi-level discourse analysis. *Computational Linguistics*, 18(4):537–544.

Moran, T., Palen, L., Harrison, S., Chiu, P., Kimber, D., Minneman, S., Van Melle, W., and Zellweger, P. (1997). I'll get that off the audio : a case study of salvaging multimedia meeting records. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 202–209. ACM Press.

Morgan, N., Baron, D., Bhagat, S., Carvey, H., Dhillon, R., Edwards, J., Gelbart, D., Janin, A., Krupski, A., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., and Wooters, C. (2003). Meetings about meetings: Research at icsi on speech in multiparty conversations. In *Proc of the ICASSP'03*.

Morris, J. and Hirst, G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17(1):21–45.

Mosvick, R. and Nelson, R. (1987). *We've got to start meeting like this! A guide to successful business meeting management*. Scott Foresman Publishers.

Murphy, K. (2002). *Dynamic Bayesian Networks: Representation, Inference and Learning*. PhD thesis, University of California, Berkely.

Murray, G., Renals, S., and Taboada, M. (2006). Prosodic correlates of rhetorical relations. In *Proceedings of HLT/NAACL ACTS Workshop*.

Nakanishi, H., Ishida, T., Ibister, K., and Nass, C. (2004). *Agent Culture: Human-Agent Interaction in a Multicultural World*, chapter 11, Designing a Social Agent for Virtual Meeting Space, pages 245–266. Lawrence Erlbaum Associates, Publishers.

Neass, A. (1966). *Communication and argument. Elements of applied semantics*. George Allen & Uwin Press.

Nemeth, C. (1983). Reflections on the dialogue between status and style: Influence processes of social control and social change. *Social Psychological Quarterly*, 46:70–74.

Newman, S. and Marshall, C. (1991). Pushing toulmin too far: Learning from an argument representation scheme. Technical Report SSL-92-45, Xerox PARC.

Newman, W. M. (1997). Better or just different? on the benefits of designing interactive systems in terms of critical parameters. In *Proceedings of the conference on Designing interactive systems: processes, practices, methods, and techniques*, pages 239–245, Amsterdam, The Netherlands.

Nguyen, T., Qui, T., Xu, K., Cheok, A., Teo, S., Zhou, Z., Mallawaarachchi, A., Lee, S., Liu, W., Teo, H., Thang, L., Li, Y., and Kato, H. (2005). Real-time 3d human capture system for mixed-reality art and entertainment. *IEEE Transactions on Visualization and Computer Graphics*, 11(6):706–721.

Niekrasz, J. and Purver, M. (2005). A multimodal discourse ontology for meeting understanding. In Bourlard, H. and Bengio, S., editors, *Proceedings of MLMI'05*. Springer-Verlag. LNCS.

Nijholt, A., Rienks, R., Zwiers, J., and Reidsma, D. (2006). Online and off-line visualization of meeting information and meeting support. *The Visual Computer*.

Nijholt, A., Rist, T., and Tuijnenbreijer, K. (2004). Lost in ambient intelligence? In *Extended abstracts on Human factors in computing systems (CHI'04)*, pages 1725–1726, Vienna, Austria.

Nijholt, A., Zwiers, J., and Peciva, J. (2005). The distributed virtual meeting room exercise. In Vinciarelli, A. and Odobez, J., editors, *Proceedings ICMI 2005 Workshop on Multimodal multiparty meeting processing, Trento, Italy*, pages 93–99.

Nishida, T. (2007). *AI for Human Computing, Lecture Notes in Artificial Intelligence 4451 (To Appear)*, chapter Social Intelligence Design. Springer-Verlag, London, UK.

Nomoto, T. and Matsumoto, Y. (1999). Learning discourse relations with active data selection. pages 158–167.

Norman, D. (2004). *Emotional Design: why we love (or hate) everyday things*. Basic Books, Cambridge, MA, U.S.A.

Nunamaker, J., Dennis, A., Valacich, J., Vogel, D., and George, J. (1991). Electronic meeting systems. *Communications of the ACM*, 34(7):40–61.

Nunamaker Jr., J., Briggs, R., and Mittleman, D. (1995). Electronic meeting systems: Ten years of lessons learned. In Coleman, D. and Khanna, R., editors, *Groupware: Technology and Applications*. Prentice Hall, Englewood Cliffs, NJ.

Nunamaker Jr., J., R.O., B., D.D., M., D.R., V., and P.A., B. (1996). Lessons from a dozen years of group support systems research: a discussion of lab and field findings. *Journal of Management Information Systems*, 13(3):163–207.

O'Connell, D., Kowal, S., and Kaltenbacher, E. (1990). Turn-taking: A critical analysis of the research tradition. *Journal of Psycholinguistic Research*, 19(6):345–373.

Ofsche, R. and Lee, M. (1981). Status, deference, influence and convenient rationalization: An application of two-process theory. Working Papers in Two-Process Theory 3, Department of Sociology, University of California.

Oh, A., Tuchinda, R., and Wu, L. (2001). Meetingmanager: A collaborative tool in the intelligent room. In *Proc. of the MIT Student Oxygen workshop*.

Oh, J. and Smith, F. (2005). Calendar assistants that learn preferences. In Saphiro, D., editor, *Persistant Assistants: Living and working with AI, workshop at the AAAI Spring Symposium 2005*. AAAI Press.

Ohsawa, Y., Matsumura, N., and Ishizuka, M. (2002). Influence diffusion model in text-based communication. In *Proc. of The eleventh world wide web conference*.

Orbons, E. (2006). Real-time 3d head reconstruction using multiple cameras. In *Proccedings of the 5th Twente Student Conference on IT*.

Orlikowski, W. and Yates, J. (1994). Genre repertoire: The structuring of communicative practices in organizations. *Administrative Science Quarterly*, 39:541–574.

Orwell, G. (1949). *1984*. Secker and Warburg.

Oviatt, S. L. (2003). *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications*, chapter 14: Multimodal interfaces, pages 286–304. Lawrence Erlbaum Associates.

Padilha, E. and Carletta, J. (2003a). Nonverbal behaviors improving a simulation of small group discussion. In *Proceedings of the 1st Nordic Symposium on Multimodal Communication*, pages 93–105.

Padilha, E. and Carletta, J. (2003b). Nonverbal behaviors improving a simulation of small group discussion. In *Proceedings of the 1st Nordic Symposium on Multimodal Communication*, pages 93–105.

Pallotta, V., Ghorbel, H., Ruch, P., and Coray, G. (2004). An argumentative annotation schema for meeting discussions. In *Proceedings of the 4th LREC conference*.

Pallotta, V., Niekrasz, J., and Purver, M. (2005). Collaborative and argumentative models of meeting discussions. In *Proceeding of CMNA-05 international workshop on Computational Models of Natural Arguments (part of IJCAI 2005)*.

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 aswering questions on meeting discussion records. In *Proceedings of workshop on Modelling Meetings Argumentation and Discourse (MMAD'06)*.

Palotta, V., Ghorbel, H., Ballim, A., Lisowska, A., and Marchand-Maillet, S. (2004). Towards meeting information systems. In Seruca, E., Filipe, J., Hammoudi, S., and Cordeiro, J., editors, *Proceedings of the ICEIS 2004: 6th International conference on enterprise information systems*, volume 4.

Panko, R. and Kinney, S. (1995). Meeting profiles: size, duration, and location. In *HICSS '95: Proceedings of the 28th Hawaii International Conference on System Sciences*, page 1002, Washington, DC, USA. IEEE Computer Society.

Pantic, M., Pentland, A., Nijholt, A., and Huang, T. (2006). Human computing and machine understanding of human behavior: A survey. In Kwek, F. and Yang, Y., editors, *ACM SIGCHI Proceedings Eighth International Conference on Multimodal Interfaces (ACM ICMI 2006)*, pages 239–248, Banff, Canada. ACM, New York.

Passonneau, R. and Litman, D. (1997). Discourse segmentation by human and automated means. *Computational Linguistics*, 23(1):103–140.

Paulsen, D. (2004). Leadership essentials: racilitation skills for improving group effectiveness. In *SIGUCCS '04: Proceedings of the 32nd annual ACM SIGUCCS conference on User services*, pages 153–160, New York, NY, USA. ACM Press.

Pelachaud, C. and Bilvi, M. (2003). Computational model of believable conversational agents. *Communication in MAS: background, current trends and future*.

Pentland, S. (2005). Socially aware computation and communication. *IEEE Computer*, pages 63–70.

Perakyla, A. (2004). Two traditions in interction research. *Britisch Journal of Social Psychology*, 43:1–20.

Perelman, C. and Olbrechts-Tyteca, L. (1969). *The New Rethoric. A tretise on Argumentation*. Notre Dame Press.

Pertaub, D., Slater, M., and Barker, C. (2002). An experiment on public speaking in response to three different types of virtual audience. *Presence: Teleoperators and Virtual Environments*, 11(1):68–78.

Phillips, J., Flynn, P. J., Scruggs, T., Bowyer, K. W., and Worek, W. (2006). Preliminary face recognition grand challenge results. In *Proceedings of the Conference on Automatic Face and Gesture Recognition 2006 (FGR'06)*, pages 15–24, Southampton, UK.

Pianesi, F., Zancanaro, M., Falcon, V., and Not, E. (2006). Toward supporting group dynamics. In Maglogiannis, I., Karpouzis, K., and Bramer, M., editors, *Proceedings of the 3rd IFIP Conference on Artificial Intelligence Applications*, pages 302–311. Springer Verlag.

Pinsonneault, A. and Kraemer, K. (1989). The impact of technological support on groups: An assessment of the empirical research. *Decision Support Systems*, 5:197–216.

Polyani, L. (1988). A formal model of the structure of discourse. *The Journal of Pragmatics*, 12:601–638.

Poole, M.S., S. D. and McPhee, R. (1985). Group decision-making as a structurational process. *Quarterly Journal of Speech*, 71:74–102.

Poppe, R., Rienks, R., and Heylen, D. (2007). Accuracy of head orientation perception in triadic situations: Experiment in a virtual environment. *Perception, to appear*. ISSN=0301-0066.

Post, W., Cremers, A., and Henkemans, O. (2004). A research environment for meeting behavior. In *Proceedings of the third workshop on Social Intelligence Design (SID)*.

Post, W., Elling, E., Cremers, A., and Kraaij, W. (2007). Experimental comparison of multimodal meeting browsers. In *To appear in Proceedings of the HCI International 2007*.

Prakken, H., Reed, C., and Walton, D. (2003). Argumentation schemes and generalisations in reasoning about evidence. In Peczenik, A., editor, *Proceedings of the 21st ivr congress*. Franz Steiner Verlag.

Purver, M., Ehlen, P., and Niekrasz, J. (2006). Shallow discourse structure for action item detection. In *Proceedings of the 2006 HLT-NAACL Workshop on Analyzing Conversations in Text and Speech*, New York City, New York, USA.

Purver, M., Niekrasz, J., and Peters, S. (2005). Ontology-based multi-party meeting understanding. In *Proceedings of CHI 2005 Workshop: CHI Virtuality 2005*.

Quinlan, J. (1993). *C4.5 : programs for machine learning*. Morgan Kaufmann, San Mateo, CA, USA.

Quirk, R. and Greenbaum, S. (1973). *A Concise Grammar of Contemporary English*. Harcourt Brace Jovanovich Inc.

Reed, C. and Rowe, G. (2001). Araucaria: Software for puzzles in argument diagramming and xml. Technical report, Department of Applied Computing, University of Dundee.

Reidsma, D. (2009). *Designing Observations of Zoiets Dergelijks*. Phd. thesis, Faculty of Creative Technologies, 3TU.

Reidsma, D., Hofs, D., and Jovanovic, N. (2005a). A presentation of a set of new annotation tools based on the nxt api. Poster at Measuring Behavior 2005. AMI-105.

Reidsma, D., op den Akker, R., Rienks, R., Poppe, R., Nijholt, A., Heylen, D., and Zwiers, J. (2005b). Virtual meeting rooms: From observation to simulation. In *Proceedings of the 4th workshop on Social Intelligence Design*.

Reidsma, D., Rienks, R., and Jovanovic, N. (2004). Meeting modelling in the context of multimodal research. In *Proc. of the Workshop on Machine Learning and Multimodal Interaction*.

Reidsma, D., van Welbergen, H., Poppe, R., Bos, P., and Nijholt, A. (2006). Towards bi-directional dancing interaction. In *International Conference on Entertainment Computing (ICEC'06)*, volume 4161 of *Lecture Notes in Computer Science*, pages 1–12, Cambridge, UK. Springer Verlag.

Reiter, S., Schuller, B., and Rigoll, G. (2006). Segmentation and recognition of meeting events using a two-layered hmm and a combined mlp-hmm approach. In *Proceedings of the International Conference on Multimedia and Expo (ICME) 2006*.

Reitter, D. (2003). Simple signals for complex rhetorics: On rhetorical analysis with rich-feature support vector models. In Seewald-Heeg, U., editor, *Sprachtechnologie für die multilinguale Kommunikation.*, St. Augustin, Germany.

Richman, L. (1987). Software systems that catch the team spirit. *Fortune*, 115(12):125–136.

Robert, H. (2000). *Roberts Rules of Order Revised*. Bartleby.com.

Rodenstein, R. and Donath, J. (2000). Talking in circles: Designing a spatially-grounded audioconferencing environment. In *Proceedings of the CHI 2000*, pages 81–88. ACM Press.

Rogelberg, S., Burnfield, J., Leach, D., and Warr, P. (2006). Not another meeting! are meeting time demands related to employee well-being. *Journal of Applied Psychology*, 91(1):86–96.

Romano Jr., N. and Nunamaker Jr., J. (2001). Meeting analysis: Findings from research and practice. In *in Proceedings of the 34th Hawaii International Conference on System Sciences*, pages 1072–1085. IEEE Press.

Rosa, E. and Mazur, A. (1979). Incipient Status in Small Groups. *Social Forces*, 58(1):18–37.

Rotaru, M. (2002). Dialog act tagging using memory-based learning. Technical report, University of Pittsburgh. Term project in Dialogue-Systems class.

Sachs, H., Schegloff, E., and Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696–735.

Sacks, H. (1992). *Lectures on Conversations, Vol 1 (Autumn 1964 Spring 1968)*. Blackwell Publishers.

Sakong, K. and Nam, T. (2006). Supporting telepresence by visual and physical cues in distributed 3d collaborative design environments. In *CHI '06 extended abstracts on Human factors in computing systems*, pages 1283–1288, New York, NY, USA. ACM Press.

Samuel, K., Carberry, S., and Vijay-Shanker, K. (1999). Automatically selecting useful phrases for dialogue act tagging. *The Computing Research Repository*.

Saurí, R., Verhagen, M., and Pustejovsky, J. (2006). Annotating and recognizing event modality in text. In *Proceedings of the 19th International FLAIRS Conference, FLAIRS 2006.*, Melbourne Beach, FL, USA.

Schank, R. and Abelson, R. (1977). *Scripts, Plans, Goals and Understanding*. Hillsdale: Lawrence Erlbaum Associates.

Schauer, H. (2000). Referential structure and coherence structure. In *Proceedings of the 7th conference on Computational Approaches to Natural Language TALN'00.*

Schauer, H. and Hahn, U. (2001). Anaphoric cues for coherence relations. In Angelova, G., Bontcheva, K., Mitkov, R., Nicolov, N., and Nikolov, N., editors, *Proceedings of the Euroconference Recent Advances in Natural Language Processing (RANLP-2001)*, pages 228–234, Tzigov, Bulgaria.

Schegloff, E. (1968). Sequencing in conversational openings. *American Anthropologist*, 70(6):1075–1095.

Schiehlen, M. (2002). Ellipsis resolution with underspecified scope. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 72–79.

Schmidt, A., Kranz, M., and Holleis, P. (2005). Interacting with the ubiquitous computer: towards embedding interaction. In *Proceedings of the joint conference on Smart objects and ambient intelligence (sOc-EUSAI'05)*, pages 147–152, Grenoble, France.

Schultz, T., Waibel, A., Bett, M., Metze, F., Pan, Y., Ries, K., Schaaf, T., Soltau, H., Westphal, M., Yu, H., and Zechner, K. (2001). The isl meeting room system. In *Proceedings of the Workshop on Hands-Free Speech Communication*.

Schum, D. and Martin, A. (1982). Formal and empirical research on cascaded inference in jurisprudence. *Law and Society Review*, 17(1):105–152.

Schwartzman, H. (1989). *The Meeting*. Plenum Press.

Scott Morton, M. S. (1971). *Management Decision Systems; Computer-based support for decision making*. Graduate School of Business Administration, Division of Research, Harvard University.

Searle, J. (1969). *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press.

Sell, J., Lovaglia, M., Mannix, E., Samuelson, C., and Wilson, R. (2004). Investigating conflict, power and status within and among groups. *Small Group Research*, 35(1):44–72.

Sellen, A. (1995). Remote conversations: The effects of mediating talk with technology. *Human-Computer Interaction*, 10(4):401–444.

Selvin, A. (2003). *Visualizing Argumentation: Software Tools for Collaborative and Educational Sense-Making*, chapter Fostering Collective Intelligence: Helping Groups Use Visualized Argumentation. Springer Verlag, London, UK.

Selvin, A., Buckingham Shum, S., Sierhuis, M., Conklin, J., Zimmermann, B., Palus, C., Drath, W., Horth, D., Domingue, J., Motta, E., and Li, G. (2001). Compendium: Making meetings into knowledge events. In *Proc. Knowledge Technologies 2001*.

Short, J., Williams, E., and Christie, B. (1976). *The social psychology of telecommunications*. John Wiley & Sons.

Shriberg, E., Dhillon, R., Bhagat, S., Ang, J., and Carvey, H. (2004). The icsi meeting recorder dialog act (mrda) corpus. In *Proc. HLT-NAACL SIGDIAL Workshop, Boston, April-May*.

Shum, D. (1994). *The evidential foundations of probabilistic Reasoning*. John Wiley & Sons.

Siegel, J., Dubrovski, V., Kiesler, S., and McGuire, T. (1986). Group processes in computer mediated communication. *Organizational Behavior and Human Decision Processes*, 37:157–187.

Sinha, P. (2002). Recognizing complex patterns. *Nature Neuroscience*, 5:1093–1097.

Slabaugh, G., Schafer, R., and Hans, M. (2002). Image-based photo hulls. Technical Report HPL-2002-28, Client and Media Systems Laboratory Hewlet Packard Laboratories, Palo Alto, CA.

Slater, M. and Steed, A. (2001). *The Social Life of Avatars: Presence and Interaction in Shared Virtual Environments*, chapter Meeting People Virtually: Experiments in Shared Virtual Environments, pages 146–171. Springer-Verlag, London, UK.

Slater, P. (1958). Contrasting correlates of group size. *Sociometry*, 21(2):129–139.

Slavin, R. (1983). When does cooperative learning increase student achievement? *Psychological Bulletin*, 94(3):429–445.

Smith, M. (1942). An approach to the study of the social act. *Psychological Review*, 49(5):422–440.

Smith, M., J.J., C., and Burkhalter, B. (2000). Conversation trees and threaded chats. In *Computer Supported Cooperative Work (CSCW'00)*, pages 97–105.

Stefik, M., Foster, G., Bobrow, D., Kahn, K., Lanning, S., and Suchman, L. (1987). Beyond the chalkboard: computer support for collaboration and problem solving in meetings. *Communications of the ACM*, 30(1):32–47.

Steidl, S., Levit, M., Batliner, A., Nöth, E., and Niemann, H. (2005). "of all things the measure is man" automatic classification of emotion and intra labeler consistency. In *ICASSP 2005, International Conference on Acoustics, Speech, and Signal Processing*.

Steiner, I. (1972). *Group process and productivity*. Academic Press.

Stolcke, A. (2002). Srilm – an extensible language modeling toolkit. pages 901–904.

Straus, S. and McGrath, J. (1994). Does the medium matter? the interaction of task type and technology on group performance and member reactions. *Journal of Applied Psychology*, 79(1):87–97.

Suthers, D. (2001). Towards a systematic study of representational guidance for collaborative learning discourse. *Journal of Universal Computer Science*, 7(3). Electronic Publication.

Suthers, D., Weiner, A., Connelly, J., and Paolucci, M. (1995). Belvedere: Engaging students in critical discussion of science and public policy issues. In *Proceedings of the the 7th World Conference on Artificial Intelligence in Education (AIED).*

Syrdal, A. K., Hirschberg, J., McGory, J., and Beckman, M. (2001). Automatic tobi prediction and alignment to speed manual labelling of prosody. *Speech communication*, 33:135–151.

Taboada, M. and Mann, W. (2006). Rhetorical structure theory: looking back and moving ahead. *Discourse Studies*, 8(3):423–459.

Tan, H. and Pentland, A. (1997). Tactual displays for wearable computing. In *Proceedings of the first International Symposium on Wearable Computing(ISCW)*, pages 84–89.

Tatum, M. (2000). Active worlds. *SIGGRAPH Computer Graphics*, 34(2):56–57.

Thomans, W. (1927). The behavior pattern and the situation. *Publications of the American Sociological Society*, 22:1–13.

Thomas, P. and Macredie, R. D. (2002). Introduction to the new usability. *ACM Transactions on Computer-Human Interaction*, 9(2):69–73.

Thorpe, E. (1998). The invention of the first wearable computer. In *Proceedings of the second International Symposium on Wearable Computing(ISCW)*, pages 4–8.

Thorpe, L. and Schmuller, A. (1954). *Contemporary Theories of Learning-with Applications to Education and Psychology.* The Ronald Press Company.

Tollmar, K. and Persson, J. (2002). Understanding remote presence. In *NordiCHI '02: Proceedings of the second Nordic conference on Human-computer interaction*, pages 41–50, New York, NY, USA. ACM Press.

Toulmin, S. (1958). *The Uses of Argument.* Cambridge University Press.

Toutanova, K., Klein, D., and Manning, C. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 173–180, Morristown, NJ, USA. Association for Computational Linguistics.

Tracy, K. and Coupland, N. (1990). Multiple goals in discourse: An overview of issues. *Journal of Language and Social Psychology*, 9:1–13.

Traum, D. (1994). *A Computational Theory of Grounding in Natural Language Conversation TR 545.* PhD thesis, Computer Science Department of the University of Rochester.

Traum, D. and Rickel, J. (2002). Embodied agents for multi-party dialogue in immersive virtual worlds. In *Proc. of the 1st Int. Joint Conf. on Autonomous Agents and Multi-agent Systems (AAMAS 2002)*, pages 766–773.

Treu, S. (1975). On-line student debate: an experiment in communication using computer networks. *International Journal of Computer and Information Sciences*, 4(1):39–51.

Tucker, S. and Whittaker, S. (2004). Accessing multimodal meeting data: Systems, problems and possibilities. In *In Proceedings of MLMI'04.* Springer-Verlag.

Tucker, S. and Whittaker, S. (2005). Reviewing multimedia meeting recordings: Current approaches. In *in Multimodal Multiparty Meeting Processing, Workshop at the 7th International Conference on Multimodal Interfaces.*

Turk, M. and Kolsch, M. (2004). *Emerging Topics in Computer Vision*, chapter Perceptual interfaces, pages 455–519. Prentice Hall.

Vallee, J., Johansen, R., , and Spangler, K. (1975). The computer conference: An altered state of communication? *The Futurist*, pages 116–121.

van Bunningen, A. H., Feng, L., and Apers, P. M. (2005). Context for Ubiquitous Data Management. In *International Workshop on Ubiquitous Data Management (UDM'05)*, pages 17–24, Tokyo, Japan.

Van Eemeren, F. (2003). A glance behind the scenes: The state of the art in the study of argumentation. *Studies in Communication Sciences*, 3(1):1–23.

Van Eemeren, F., Grootendorst, R., and Kruiger, T. (1987). *Handbook of Argumentation Theory*. Foris publications.

Van Eemeren, F., Grootendorst, R., and Snoeck Henkemans, F. (2002). *Argumentation*. Lawrence Erlbaum Associates.

van Gelder, T. (2002). Argument mapping with reason!able. The American Philosophical Association Newsletter on Philosophy and Computers.

Van Gelder, T. (2003). *Visualizing Argumentation: Software Tools for Collaborative and Educational Sense-Making*, chapter Enhancing Deliberation through Computer-Supported Argument Visualization. Springer Verlag, London, UK.

Van Vree, W. (1999). *Meetings, Manners and Civilisations*. Leicester University Press.

Veerman, A. (2000). *Computer-supported collaborative learning through argumentation*. PhD thesis, University of Utrecht.

Verbree, A., Rienks, R., and Heylen, D. (2006). Dialogue-act tagging using smart feature selection: results on multiple corpora. In *The first International IEEE Workshop on Spoken Language Technology (SLT)*, Palm Beach, Aruba.

Verbree, D. (2006). On the structuring of discourse based on utterances automatically classified. Master's thesis, University of Twente, Human Media Interaction Group, Faculty of EEMCS.

Vertegaal, R. (1998). *Look who is talking to whom*. PhD thesis, University of Twente.

Vissers, G., Heyne, G., Peters, V., and Guerts, J. (2001). The validity of laboratory research in social and behavioral science. *Quality & Quantity*, 35:129–145.

Vreede de, G. and Muller, P. (1997). Why some gss meetings just don't work: Exploring success factors of electronic meetings. In *Proceedings of the 7th European Conference on Information Systems (ECIS)*, pages 1266–1285.

Wainer, J. and Braga, D. (2001). Symgroup: applying social agents in a group interaction system. In *Proceedings of the 2001 International ACM SIGGROUP Conference on Supporting Group Work*, pages 224–231. ACM Press.

Wainfan, L. and Davis, P. (2004). Challenges in virtual collaboration. video conferencing, audio conferencing and computer mediated communications. Technical Report MG-273, RAND Corporation, National Defence Research Institute. U.S.A., Santa Monica, U.S.A.

WainHouse-Research (2006). Rich media conferencing forecast update. The Wainhouse Research Bullitin, Volume 7, Issue 41.

Walther, J., Anderson, J., and Park, D. (1994). Interpersonal effects in computer-mediated interaction: A meta analysis of social and antisocial communication. *Communication Research*, 21(4):460–487.

Walton, D. (1996). *Argument Structure, A pragmatic Theory*. University of Toronto Press.

Walton, D. and Reed, C. (2003). *Anyone Who Has a View: Theoretical Contributions to the Study of Argumentation*, chapter Diagramming, Argumentation Schemes and Critical Questions, pages 195–211. Kluwer, Dordrecht, The Netherlands.

Wang, J. (2006). Questions and the exercise of power. *Discourse and Society*, 17(4):529–548.

Webber, B., Knott, A., and Joshi, A. (1999). Multiple discourse connectives in a lexicalized grammar for discourse. In *Proceedings of the Third International Workshop on Computational Semantics*.

Weisenbaum, J. (1966). Eliza: A computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1).

Weiser, M. (1991). The computer of the 21st century. *Scientific American*, 265(3):66–75.

Weldon, E. and Weingart, L. R. (1993). Group goals and group performance. *British Journal of Social Psychology*, 32:307–334.

Wellner, B., Pustejovsky, J., Havasi, C., Rumshisky, A., and Saurí, R. (2006). Classification of discourse coherence relations: An exploratory study using multiple knowledge sources. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, pages 117–125, Sydney, Australia. Association for Computational Linguistics.

Wellner, P., Flynn, M., and Guillemot, M. (2004). Browsing recorded meetings with ferret. In *In Proceedings of MLMI'04*. Springer-Verlag.

West, C. and Zimmerman, D. (1983). *Small Insults: A study of interruptions in cross-sex conversations between unaquainted persons*, chapter in Language, Gender and Society, pages 103–117. Newbury House.

West, M. (2002). Sparkling fountains or stagnant ponds: An integrative model of creativity and innovation implementation in work groups. *Applied Psychology*, 51:355–424.

West, M. (2003). *Effective Teamwork*. British Psychological Society.

West, M., Tjsovold, D., and Smith, K. (2003). *The international handbook of organizational temawork and cooperative working*. John Wiley and Sons.

Whittaker, S. (2002). *The Handbook of Discourse Processes*, chapter Theories and Methods in Mediated Communication, pages 243–286. Erlbaum, New York NJ. U.S.A.

Whittaker, S. (2005). User requirements in ami. In *Symposium 'annotating and measuring meeting behavior' at Measuring Behavior 2005*.

Whittaker, S. and O'Conaill, B. (1993). An evaluation of video mediated communication. In *CHI '93: INTERACT '93 and CHI '93 conference companion on Human factors in computing systems*, pages 73–74, New York, NY, USA. ACM Press.

Whittaker, S., Terveen, L., and Nardi, B. (2000). Let's stop pushing the envelope and start addressing it: a reference task agenda for HCI. *Human Computer Interaction*, 15(2-3):75–106.

Wigmore, J. (1931). *The principles of Judicial Proof, 2nd ed.* Little, Brown and Company.

Willard, C. (1976). On the utility of descriptive diagrams for the analysis and criticism of arguments. *Communication Monographs*, 43:308–319.

Willard, D. and Strodbeck, F. (1972). Latency of verbal response and participation in small groups. *Sociometry*, 35:161–175.

Witten, I. H. and Frank, E. (2000). *Data Mining: Practical machine learning tools with Java implementations*. Morgan Kaufmann.

Wolf, F. and Gibson, E. (2005). Representing discourse coherence: A corpus-based study. *Computational Linguistuistics*, 31(2):249–288.

Wooldridge, M. and Jennings, N. (1995). Intelligent Agents: Theory and Practice. *Knowledge Engineering Review*, 10(2):115–152.

Wrede, B. and Shriberg, E. (2003a). Hotspots in meetings: Human judgements and prosodic cues. In *Proceedings of the European Conference on Speech Communication and Technology EUROSPEECH*.

Wrede, B. and Shriberg, E. (2003b). The relationship between dialogue acts and hot spots in meetings. In *Proceedings of the IEEE Speech Recognition and Understanding Workshop location = St. Thomas, Virgin Islands*.

Wynn, E. (1979). *Office conversation as an Information Medium.* PhD thesis, University of California, Berkely. Unpublished.

Xiao, J., Stasko, J., and Catrambone, R. (2002). Embodied conversational agents as a ui paradigm: A framework for evaluation. In *Proc. of workshop Embodied conversational agents - let's specify and evaluate them! at AAMAS 2002*.

Yngve, V. (1970). On getting a word in edgewise. In *Papers from the Sixth Regional Meeting of the Chicago Linguistic Society*, pages 567–577.

Yoshimi, J. (2004). The structure of debate. Technical report, University of Claifornia, Merced.

Zander, A. (1979). The psychology of group processes. *Annual Review of Psychology*, 30:417–451.

Zechner, K. (2002). Automatic summarization of open-domain multiparty dialogues in diverse genres. *Computational Linguistics*, 28(4):447–485.

Zhang, D., Gatica-Perez, D., Bengio, S., and McCowan, I. (2006). Modeling individual and group actions in meetings with layered hmms. *IEEE Transactions on Multimedia,*, 8(3):509–523.

Zhang, D., Gatica-Perez, D., Bengio, S., and Roy, D. (2005). Learning influence among interacting markov chains. *Advances in Neural Information Processing Systems (NIPS)*, 18.

Zimmerman, M., Dielmann, A., and Shriberg, E. (2006). Joint segmentation and classification of dialogue acts in multi-party meetings. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.