Cognitive Architecture of Multimodal Multidimensional Dialogue Management

DISSERTATION

zur Erlangung des Grades
des Doktors der Ingenieurwissenschaften (Dr.-Ing.)
der Naturwissenschaftlich-Technischen Fakultät
der Universität des Saarlandes

vorgelegt von
Andrei Valeryavich Malchanau

Saarbrücken, 2019

Dekan der Naturwissenschaftlich-Technische Fakultät: Prof. Dr. rer. nat. Guido Kickelbick

Mitglieder des Prüfungsausschusses:

Vorsitzender: Prof. Dr.-Ing. Georg Frey

Berichterstattender: Prof. Dr. Dietrich Klakow

Berichterstattender: Prof. Dr. Harry Bunt

Beaufsichtiger: Dr. Mohammad Molayem

Tag der mündlichen Prüfung: 19.02.2019

Eidesstattliche Versicherung

Hiermit versichere ich an Eides statt, dass ich die vorliegende Arbeit selbstständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus anderen Quellen oder indirekt übernommenen Daten und Konzepte sind unter Angabe der Quelle gekennzeichnet. Die Arbeit wurde bisher weder im In- noch im Ausland in gleicher oder ähnlicher Form in einem Verfahren zur Erlangung eines akademischen Grades vorgelegt.

Ort, Datum	(Unterschrift)

Abstract

Numerous studies show that participants of real-life dialogues happen to get involved in rather dynamic non-sequential interactions. This challenges the dialogue system designs based on a reactive interlocutor paradigm and calls for dialog systems that can be characterised as a proactive learner, accomplished multitasking planner and adaptive decision maker.

Addressing this call, the thesis brings innovative integration of cognitive models into the human-computer dialogue systems. This work utilises recent advances in Instance-Based Learning of Theory of Mind skills and the established Cognitive Task Analysis and ACT-R models. Cognitive Task Agents, producing detailed simulation of human learning, prediction, adaption and decision making, are integrated in the multi-agent Dialogue Manager. The manager operates on the multidimensional information state enriched with representations based on domain- and modality-specific semantics and performs context-driven dialogue acts interpretation and generation. The flexible technical framework for modular distributed dialogue system integration is designed and tested. The implemented multitasking Interactive Cognitive Tutor is evaluated as showing human-like proactive and adaptive behaviour in setting goals, choosing appropriate strategies and monitoring processes across contexts, and encouraging the user exhibit similar metacognitive competences.

Zusammenfassung

Zahlreiche Studien zeigen, dass reale Dialoge sich oft durch dynamische nicht-sequentielle Interaktionen auszeichnen. Dies stellt Dialogsystemsdesigns in Frage, die auf einem reaktiven Gesprächspartner-Paradigma basieren und erfordert Dialogsysteme, die zu einem proaktiven Lerner, multitaskingfähigen Planer und zu einem adaptiven Entscheidungsträger werden.

In dieser Arbeit werden auf innovative Weise kognitive Modelle in das Mensch-Computer-Dialogsystem integriert. Hierbei werden die neuesten Fortschritte im Bereich des 'Instance-Based Learning of Theory of Mind Skills' und die etablierten Cognitive Task Analysis und ACT-R Modelle verwendet. Cognitive Task Agents, die Details des menschlichen Lernens, der Anpassung und der Entscheidungsfindung simulieren, sind in den Multi-Agenten Dialogmanager integriert. Der Manager arbeitet mit einem multidimensionalen Informationszustand, der mit den fusionierten Repräsentationen der domänen- und modalitätsspezifischen Semantik angereichert ist und fhrt eine kontextgesteuerte Dialogaktualisierung und -generierung durch. Eine Plattform zur modularen Integration verteilter Dialogsysteme ist entwickelt und getestet. Der implementierte multitaskingfähige Interaktive Kognitive Tutor wird evaluiert: er zeigt ein menschenähnliches proaktives Verhalten beim Festlegen von Zielen, Auswählen geeigneter Strategien und beim Überwachen von Prozessen ber Kontexte hinweg und unterstützt vergleichbare metakognitive Kompetenzen des Nutzers.

Acknowledgments

I would like to thank many people who made the time during my PhD project stimulating and enjoyable. This thesis could not have been written without the help and inspiration of many people around me.

First of all, I would like to thank my supervisor Prof. Dietrich Klakow for the unique opportunity given to me to perform this research. I am grateful to have found an intellectual and creative environment at the Spoken Language System Group of the University of Saarland - you, Dietrich, created! I would like to thank all my colleagues, especially Dr Volha Petukhova, for co-authoring most of my recent papers, creative discussions and timely support.

My word of special gratitude also goes to Prof Harry Bunt for his enthusiasm and interest in my work, and encouragement that helped me so much to bring to this point - PhD defence! Thank you for frequent inspiring discussions that enhanced my professional competence and broaden my knowledge.

The research reported in this thesis has been partially performed within the FP7 EU-funded project METALOGUE , under grant reference 611073. My thanks go to the members of the Metalogue consortium: Jan Alexandersson, the whole DFKI coordinating and research team, Nick Campbell and Saturnino Luz (Trinity College Dublin), Alexander Stricker and Christian Dold (Charamel GmbH), Dimitris Koryzis (Hellenic Parliament), Peter van Rosmalen, Dirk Börner and Jan Schneider (Open University of the Netherlands), Joy van Helvert, Michael Gardner and Emmanuel Ferreyra (University of Essex). I specially would like to thank the Metalogue team from the University of Groningen: Prof Niels Taatgen, Christopher Stevens, Harmen de Weerd and Fokie Cnossen. It was very inspiring and constructive collaboration which contributed to this thesis the most, thanks to their support in cognitive modelling tasks.

At last, but certainly not at least, I would like to thank my family and friends for support and faith in me.

Saarbrücken,
January, 2019

Andrei Malchanau

Contents

Chapter 1

Introduction

1.1 Motivation

The increasing complexity of human-computer systems and interfaces results in an increasing demand for intelligent interaction that is natural to users and that exploits the full potential of spoken and multimodal communication. Much of the research in human-computer system design has been conducted in the area of task-oriented systems, especially for information-seeking dialogues concerning well-defined tasks in restricted domains – see Table 1.1 for the main paradigms used for dialogue modelling in domains of varying complexity.

Many existing dialogue systems represent a set, often very rigid, of possible dialogue state transitions for a given dialogue task. Dialogue states are typically defined in terms of dialogue actions, e.g. question, reply, inform, and slot filling goals. States in a finite state transition network are often used to represent the dialogue states (Bilange, 1991; Dahlbäck and Jönsson, 1998). Some flexibility has been achieved when applying statistical machine learning methods to dialogue state tracking (Williams et al., 2013). Statistical dialogue managers were initially based on Markov Decision Processes (Young, 2000) where given a number of observed dialogue events (often dialogue acts), the next event is predicted from the probability distribution of the events which have followed these observed events in the past. Partially Observable Markov Decision Processes (Williams and Young, 2007) model unknown user goals by an unknown probabilistic distribution over the user states. This approach is considered as the state-of-the-art in task-oriented spoken dialogue systems, see [Young et al., 2013]. Recently, deep neural networks have gained a lot of attention (Henderson et al, 2013; 2014). Hierarchical recurrent neural networks and memory networks have been proposed to generate open domain dialogues and build end-to-end dialogue systems trained on large amounts of data without any detailed specification of information states (Serban et a., 2016; Sukhbaatar et al., 2015). The real challenge for end-to-end frameworks is the decision-taking problem related to the dialogue management for goal-oriented dialogues. Statistical and end-to-end approaches require really large amounts of data while offering a rather limited set of dialogue actions (Kim et al., 2015; Henderson et al., 2008).

Technique	Example task	Dialogue phenomena handled
Finite state script	Long-distance calling	User answers questions
Frame based	Getting train timetable information	User asks questions, simple clarifications by the system
Information State Update	Travel booking agent	Flexible shifts between pre-determined topics/tasks Refined grounding mechanisms
Plan based	Kitchen design consultant	Dynamically generated topic structures, e.g. negotiation dialogues
Agent based	Disaster relief management	Different modalities, e.g. planned world and actual world Collaborative planning and acting
Probabilistic approaches	Various information-seeking tasks, negotiation games	Dialogue policies design, i.e. learning combined with the most approaches mentioned above
Chat-oriented; interactive pattern matching /template-based	Retail 'chat commerce' Psychotherapies, personal assistant	Question-answering skills Social interactive aspects

Table 1.1: State-of-the-art techniques for task-oriented dialogue system.

While statistical dialogue systems may perform well on simple information-transfer tasks and end-to-end approaches handle well chatbot conversations, they are mostly unable to manage real-life communication in complex settings like, for example, multi-party conversations, tutoring sessions and debates. More conversationally plausible dialogue models are based on rich representations of dialogue context for flexible dialogue management, e.g. information-state updates (ISU, Traum et al., 1999; Bunt, 1999; Bos et al., 2003; Keizer et al., 2011). Other approaches to dialogue processing and management are built as full models of rational agency accounting for planning and plan recognition (Cohen and Perrault, 1979; Carberry, 1990; Sadek, 1991). Plan construction and inference are activities that can however easily get very complex and become computationally intractable. Alternatively, dialogue plans and strategies can be learned and adapted through reinforcement learning, [Sutton and Barto, 1998].

The research community is currently targeting more flexible, adaptable, open-domain multimodal dialogue systems. Advances are made in modelling and managing multi-party interactions, e.g. for meetings or multi-player games, where approaches developed for two-party dialogue are extended in order to model phenomena specific to multi-party interactions. Nevertheless, simple command/control and query/reply systems prevail. Some dialogue systems developed for research purposes allow for more natural conversations, but they are often limited to a narrow manually crafted domain and to rather restricted communication behaviour models, e.g. often modelled on information retrieval tasks. In some cases, these restrictions are imposed deliberately by the researchers to be able to investigate a limited set of dialogue phenomena without having to deal with unrelated details. However, this reduces the practical realism of the dialogue system.

Expectations of the users of today are rather high and require a real-time engagement with highly relevant personalised content that mimics human natural behaviour and is able to adapt to changing users' needs and goals. Nowadays, there is a growing interest in Artificial Intelligence (AI)-powered conversational systems that are able to learn and reason, to facilitate realistic interactive scenarios with realistic assets and lifelike, believable characters and interactions. AI models may represent rather complex research objects. Despite their acknowledged potential, generating plausible AI models from scratch is challenging.

For instance, cognitive models were successfully integrated into intelligent tutoring and intelligent narrative systems, see [Paiva et al., 2004, Riedl and Stern, 2006, Van Lehn, 2006, Ritter et al., 2007, Lim et al., 2012]. Since such models produce detailed simulations of human performance encompassing many domains such as learning, multitasking, decision making, and problem solving, they are also perfectly capable to play the role of a believable human-like agent in various human-agent settings. Although the abilities of cognitive agents continue to improve, human-agent interaction is often awkward and unnatural. The agents most of the time cannot deliver human-like interactive behaviour, but deal well with task-related actions thanks to the use of well-defined computational cognitive task models.

This thesis presents an approach to the incorporation of cognitive task models into Information State Update (ISU) dialogue management as a part of a multimodal dialogue system. Such integration has important advantages. The ISU methodology has been applied successfully to a large variety of interactive tasks, e.g. information seeking (Keizer et al., 2011), human-robot communication (Peltason and Wrede, 2011), instruction giving (Lauria et al., 2001), and controlling smart home environments (Bos et al, 2003). Several ISU development environments are available, such as TrindiKit (Larsson and Traum, 2000) and Dipper (Bos et al., 2003). The ISU approach provides a flexible computational model for understanding and generation of dialogue contributions in terms of effects on the information states of the dialogue participants. ISU models account for the creation of (shared) beliefs and mechanisms for their transfer, and have well-defined machinery for tracking, understanding and generation of natural human dialogue behaviour.

Cognitive modelling of human intelligent behaviour, on the other hand, enables deep understanding of complex mental task processes related to human comprehension, prediction, learning and decision making. Threaded cognition (Salvucci and Taatgen, 2008) and Instance-Based Learning (Gonzalez and Lebiere, 2005) models developed within the ACT-R cognitive architecture (Anderson, 2007) are used to design a cognitive agent that can respond and adapt to new situations, in particular to a communicative partner changing task goals and strategies. The designed cognitive task agents are equipped with Theory of Mind skills (Premack and Woodruff, 1978) and are able to use task knowledge not only to determine their own actions, but also to interpret the human partner's actions, and to adjust their behaviour to whom they interact with. In this way, flexible adaptive dialogue system behaviour was achieved in dynamic non-sequential interactions. The integrated cognitive agents do not only compute the most plausible task action(-s) given their understanding of the partner's actions and strategies, provide alternatives and plans possible outcomes, and therefore are able to adapt their behaviour to their partners. They also know why they select a certain action and can explain why the choices made lead to what specific outcome. This enables the agent to act as a cognitive tutor, supporting the development of the (meta)cognitive skills of a human learner. Finally, the task agent can be built using rather limited real or simulated dialogue data: it is supplied with initial state-action templates encoding domain knowledge and the agent's preferences, and the agent further learns from the collected interactive experiences.

The presented study investigates the core properties of cognitive models that underlie human task-related and interactive dialogue behaviour, shows how such models provide a

basis for dialogue management and can be integrated into a dialogue system, and assesses the resulting system usability. As the use and evaluation cases, our simulated agents and human actors participated in (meta)cognitive skills training within debate and negotiation based scenarios.

The proposed dialogue manager architecture incorporates cognitive task agents (different for different interactive tasks). While the ISU modelling approach adopted in this work operates with higher-level concepts (mental attitudes), cognitive models operate on the low-level concepts and principles described in cognitive architectures (e.g. memory chunks in the applied ACT-R). Cognitive models specify human cognitive processes based on in-depth analysis of the functioning of the human brain from a biological or neurological point of view. Both approaches complement each other resulting in an adequate and flexible computational model of complex multimodal dialogue behaviour.

1.2 Research questions

International dialogue modelling research has so far produced great rule-based and statistical multimodal dialogue systems capable of interacting with structured data bases, e.g., time tables and restaurants. Such systems typically exhibit reactive behaviour and are largely scripted/hard-coded. A single or pre-defined strategy is pursued. Systems also lack the capability reflecting about and regulating its interactive and task-related behaviour.

If a computer system that is engaged in conversation with a human can act proactively in the generation of hypotheses about upcoming utterances and dialogue acts, it is prepared for the incoming utterance in real time and can switch rapidly between prepared strategies for continuing the dialogue. A system that possesses and applies knowledge about common forms of interaction to make inferences about its users' behaviour, can produce more adequate human-like interactive behaviour.

The main reasons why people can communicate effectively and efficiently and systems cannot are the following:

1. computer dialogue systems do not have the rich experience and background knowledge that human participants have;
2. humans are able to process and perform several actions (both task-related and communicative ones) simultaneously, while dialogue systems largely are not, and if they do it mostly happens by accident rather than by design;
3. human dialogue participants are able to monitor, assess and reason about their own and their partner's progress and systems are not.

When developing a dialogue system that is able to communicate with its users efficiently and in a natural way, the Dialogue Manager (DM) as a central dialogue system component needs to be designed enabling

1. rich experience and background knowledge to be incorporated and efficiently used;

2. management of multi-tasking multimodal interactive behaviour based on a flexible and accurate computational model of such behaviour;
3. motivated extensions and machinery to handle various natural interaction complexities;
4. smooth and robust processing.

In order to comply with these requirements, several questions have to be addressed.

The main goal of the research presented here is the realization of flexible and adaptable multimodal dialogue management driven by cognitive modelling of human dialogue behaviour. Adaptable means being flexible to address and capture individual and modality-depended differences. An important action to achieve effective results, is to improve the system's knowledge and its ability to understand, control and manipulate their own and the user's cognitive processes, i.e. provide *proactive control* over the interaction.

First, the research concentrates on cognitive aspects of dialogue modelling as well as on adaptive strategies of dialogue management. To be able to exhibit adaptive behaviour, the system is required to monitor its own running processes, to connect and organise different types of information, test and modify, predict, and consequently plan and reason about future actions, and perform all these tasks based on the system's understanding of the partner's behaviour of the same kind. Such processes, referred to as metacognition, play an important role in guiding and regulating human intelligent behaviour, e.g. monitoring actions, assessing the degree to which dialogue participants understand their own and others' behaviour, obtain and apply new information, recognise failures, employ effective repair strategies and adapt their behaviour (reactively and pro-actively) to the performance and needs of the others. Adaptability and pro-activity are often related to human cognitive capabilities of monitoring, reflection and regulation.

Another function of metacognition is to improve learning in the primary cognitive process. The following is a simple example of such an improvement based on rehearsal. Memory traces that reoccur often are strengthened. However, depending on a task we may decide that we need to memorise a particular piece of information stronger. For instance, a pin-code of a bank card occurs only once in the letter that the bank sent to you. To make sure the code is retained in your memory the metacognitive strategy of rehearsal may be employed. Indeed, developmental studies have shown, rehearsal is a learned strategy: very young children do not use rehearsal at all, and slightly older children only rehearse the last item they encountered.

Yet another function of metacognition concerns reasoning about other people's intentions and knowledge. Learning reasoning strategies, Theory of Mind skills, is important in many domains, e.g. the use of language [Van Rij et al., 2010] and in playing knowledge games [Meijering et al., 2012]. Such skills develop relatively late in children, and are in some cases hard for adults. A more elaborate form of these reasoning skills is important in dialogue interaction.

Second, for a dialogue system to be able to use and understand multimodal dialogue utterances, it has to recognise the communicative functions of utterances (multiple and complex intentions) in context. Since computer dialogue systems do not possess the rich

experience and background knowledge that human participants have, they need to learn. We have observed in the past how experts and dialogue system designers successfully pre-programmed (coded) dialogue system behaviour. Considerable efforts have been undertaken to create large knowledge bases, and also to enhance access to them. Types of explicit learning have been explored, mainly exploited by the robotics community, like learning by instructions [Crangle and Suppes, 1994] and imitation learning [Seabra Lopes and Teixeira, 2000] where the machine is asked to repeat certain actions under certain context conditions. A broad range of data-driven systems that operate on the basis of learnable features are designed. Several machine-learning approaches are successfully used for this purpose.

Although building various types of classifiers as such is out of the scope of this thesis, detailed specification and formalisation of the updates they trigger in the dialogue management context model is one of the most important prerequisites when designing a dialogue system and will be discussed here in depth. The formulation of an update semantics for multimodal multifunctional dialogue utterances calls not only for the further specification of dialogue acts as update operators building up on previous related work of Petukhova (2011) and Bunt (2014), but also for the detailed specification of the semantic content addressed in the dialogue contributions. Semantic content is often domain-dependent, which is at least true in case of task-related domain-specific acts, and needs to be computed using multiple information sources like input from recording devices, synchronised and processed multimodal data streams, syntactic and semantic representations obtained, external knowledge repositories and databases, contextual information available in the dialogue history, and structural information reconstructed from various dependence and discourse relations between various dialogue units. This list would not be complete without the detailed analysis of the task(-s), constraints associated with their execution and validation criteria, e.g. in our application - debate and negotiation skills training goals.

Third, user cognitive state-aware dialogue management strategies need to be implemented. A multi-agent dialogue manager operating on the basis of an articulate context model enables multiple simultaneous and independent updates, including update mechanisms describing how a participant's context model may change during a dialogue. Fundamental principles governing human communication such as rationality, cooperation and ethics are related to general cognitive processes. We establish connections between the cognitive models and the interaction models, and specify overall mechanisms that underlie system communication strategies dependent on information about the current state of the task, with multiple and dynamically changing goals and interactive situations. The specified uniform information state consists of mental representations of participants' beliefs and attitudes related to participants' multimodal behaviour, underlying tasks and their progress, processing successes and failures, perception of the environment and how participants are situated in it, and to participants' rights and obligations given the commonly (or culturally) accepted norms and conventions for pleasant and successful interaction. The model accounts for complexities of natural human communicative behaviour such as multidimensionality (multitasking) and multimodality.

Fourth, the proposed approach needs to be evaluated both in terms of technical performance and user acceptance. In addition, a variety of applications can be employed to

show the viability of an approach. For this purpose, two different interactive applications are evaluated - Virtual Debate and Virtual Negotiation Coaches. Typically, dialogue systems are evaluated based on the users' perception of the usefulness of the system and their satisfaction with the way the task was completed. For this, we have developed satisfaction questionnaires which are filled in by the users after completing a tutoring session with the dialogue system. Many parameters are taken into account in the questionnaires to elaborate the conclusions on the system performance. We also define objective quality measures that can be automatically derived from test interactions, e.g. log files. Additionally, tools are designed to enable expert evaluations, e.g. to perform conformity checking and refinements, but also for human experts to be able to overwrite/cancel system decisions. We implemented several types of visualisation of the interactive and (meta-)cognitive processes.

To sum up, the main research questions are:

1. What are the core properties of cognitive models of metacognitive processes as a basis for a dialogue management and how can they be incorporated into a dialogue system?

2. What are advanced knowledge-based and data-driven techniques to (1) obtain rich experience and background knowledge; (2) enable parallel processing of information from multiple sources and (3) propose adaptive management?

3. How to represent information in an uniform way to avoid unnecessary overheads for "translation" between system components allowing for re-use of inference mechanisms at different stages of processing?

4. How to assess the system's performance in terms of (1) effective and adequate processing, (2) adaptive strategies in management and generation, (3) overall robustness and efficiency, and (4) user satisfaction?

1.3 Approach and starting points

The design of a dialogue system that can be engaged in complex multimodal interaction showing intelligent human-like behaviour may be expected to benefit from a good understanding of human dialogue behaviour and from the incorporation of mechanisms that are important for human dialogue communication. To build in 'intelligence' into the system, often means to equip the system with (meta-)cognitive skills: to monitor each other's interactive behaviour, to make use of resources and strategies available, to connect and organise different types of information, test and modify, predict, and consequently plan and reason about future actions. A good way to provide a dialogue system with (meta-)cognitive abilities is to understand how people acquire such skills while training them. To enable fundamentally deeper understanding of metacognitive processes and the nature of the acquisition of such skills, the system should share and vary responsibilities in observing, monitoring, experiencing and executing different tasks in multiple contexts enabling different task-related and interactive strategies. As a theoretical basis for developing such an account, the ACT-R cognitive architecture is used (Anderson, 2007). ACT-R provides a simulation

system for general cognitive processes. In the last decade considerable progress has been made in building plausible models of human intelligent behaviour, encompassing multiple domains in cognition, including acquisition of skills, memory, attention, decision making, multitasking, user modelling (Nijboer et al.,2016; Gunzelmann et al., 2009; Salvucci, 2001; Altmann and Gray, 2008; Ritter et al., 2007) and metacognition (Stevens et al., 2016). The main purpose of these models has been to further develop theory within cognitive science. Based on the ACT-R cognitive architecture and its extension of Primitive Information Processing Elements (PRIMs) theory (Taatgen, 2013), cognitive agents are built that are able to mimick human behaviour in many different domains. A key capability of the agents is that they have metacognitive abilities: they can respond and adapt themselves to the user and ensure proactive cognitive control of its own and the user's interactive actions. The main input for the agent is in the form of prior experiences that the agent uses for future decisions (Instance-Based Learning, IBL). Along with using experiences to determine its own decisions, the agent uses them to interpret behaviour of the others (i.e. human partners). In this way, the agent is able to adapt its behaviour to the other participants. Thus, agents build representations of the partners they interact with, and modify their own behavior accordingly. The agent's behaviour evolves over time: the agent gains additional instances (through experience), or evaluates the success rate of its current instances.

Human dialogue behavior is governed not only by cognitive processes of how people perceive, process, store, and apply information about other people in interactive situations, but also by fundamental principles of human social interaction such as cooperativity, rationality, sociality and ethics. These principles have been discussed at some length in the literature (e.g. Grice, 1975; Allwood et al., 2000; Bunt, 2000a). Formal computational multidimensional models, that are specific enough to use these principles to guide the interactive behaviour of a dialogue system, have been built that exploit the multifunctionality of dialogue contributions, see Bunt, 1999, 2007 and 2014; Keizer et al., 2011; Petukhova, 2011.

The methodology for modelling agents' interactive behavior is in the first place that of advancing a computational multidimensional model of multimodal interactive human behavior and its context(s). For human-agent interaction management design, we adopted an information-state or context-change approach (Poesio and Traum, 1998; Bunt, 2000a and 2000b; Traum and Larsson, 2000), which analyses dialogue utterances in terms of their effects on the dialogue context or 'information state'. In particular, we use the theoretical framework of Dynamic Interpretation Theory (DIT) for its precise definitions of communicative functions and dialogue context.

Following this approach, dialogue context and dialogue acts are the main ingredients of the dialogue model. In DIT, dialogue context is understood as the totality of conditions that influence the generation and understanding of communicative behaviour. This includes information about (1) the participants' information about the underlying task and its domain, as well as their beliefs about the dialogue partner's information of this kind (semantic context); (2) the participant's state of processing and model of the partner's state of processing (cognitive context); (3) the availability and properties of communicative and perceptual channels, and the partner's presence and attention (physical/perceptual context);

(4) communicative obligations and constraints (social context); (5) the preceding dialogue contributions ('dialogue history') and possible discourse plans (linguistic context). The dialogue context is partly dynamic, in the sense of changing during a dialogue as the result of the participants interpreting each other's communicative behaviour, reasoning with the outcomes of these processes, and planning further activities. Since these changes are essential in determining the continuation of the dialogue, we study them in detail in terms of dialogue acts.

Dialogue acts are defined as operators that update contexts in certain ways, which can be described by the communicative function and the semantic content of that dialogue act. The semantic content (propositional, referential) corresponds to what the utterance is about (what objects, events, etc., does it refer to; what propositions involving these elements are considered). In DIT, communicative functions are defined as specifications of the way semantic content is to be used by the dialogue partner to update his information state when he understands the utterance correctly.

A key methodology to be used is data-driven system design. We base our analysis and system components development on the standard well-specified and evaluated ISO 24617-2 data model. The data model contains formalised descriptions of the data objects involved, and specifies relations between them. Contents of the model are formally represented by means of typed feature structures and represented in XML-based Dialogue Act MarkUp Language (DiAML). This does not only capture the structure and relations in diverse types of data and linguistic annotations, but also facilitates the exchange of information between different processing modules of the developed system.

The cognitive, learning and interaction models are integrated to direct the operation of a dialogue manager module. The dialogue manager acts over a shared information state that incorporates and manipulates information from all models. Goals and sub-goals have the form of tasks that, in the debate setting, resulted from the detailed cognitive task analysis associated with the training goals and, in multi-issue bargaining, are generated by the cognitive model that uses the Instance-Based Learning approach. Cognitive Task Agents operate on the current situational context and on the assumption and expectations how the task should progress. The Dialogue Manager devises intentions and system-specific goals, and dynamically plans how the intentions can be achieved by taking into account the available interaction patterns and conditions/requirements arising from the context.

Putting it all together, we propose a *multidimensional model of multimodal dialogue interaction* which incorporates task related cognitive models. The Multi-Agent Dialogue Manager we designed, operates on the basis of this model. We propose the novel methodology of integrating *Cognitive Task Agents* into the dialogue system, providing models that become system components. Based on an understanding of metacognitive processes and the nature of metacognitive skills, the system has varied responsibilities in observing, monitoring, experiencing and executing different tasks (*multiperspective* dialogue), is able to play multiple roles in dialogue, some of them simultaneously. Finally, the implementation of the designed models enables the evaluation of the scientific accomplishments of this thesis and opens perspectives to further improve the quality of human-computer communication.

1.4 Contributions of this thesis

The thesis is targeting advancements in: (1) integration of cognitive, interaction and learning models into a single overarching dialogue modelling paradigm; (2) data-driven dialogue system design methodology; (3) incorporation of cognitive task agents into multi-agent dialogue management to achieve adaptive and proactive multiperspective system behaviour; (4) the design of a flexible technical integration framework for a modular distributed dialogue system incorporating state-of-the-art components; and (5) multimodal dialogue system evaluation, applying usability metrics defined in ISO standards to assess both technical performance and user acceptance.

Comprehensive cognitive interaction models have not been realised before but promise significant improvement for realism, flexibility and the ability to provide an engaging experience in interactive tutoring. Previous research showed that successful forms of learning and teaching are organised as interactive social processes (Bereiter, 2005). The effectiveness of interactive learning strongly depends on the quality of the interaction. We designed a multimodal multidimensional dialogue model that captures multitasking adaptive interactive learning behaviour accurately. The model is incorporated into and runs as part of the Dialogue Manager of the interactive tutoring system used to train metacognitive skills. The model accounts for a variety of human communicative interactive capabilities such as information exchange, monitoring information processing and application, evaluation of grounding processes and mutual understanding, manage the use of time, taking turns, and monitor contact and attention. The model is also built on a sound comprehensive model of metacognitive processes concerning human decision taking based on monitoring, reflection and regulation of speaker's own and partners' behaviour (Theory of Mind).

Secondly, we address a steadily growing interest in data-driven modelling of phenomena related to natural multimodal interactive processes. Massive amounts of data, including dialogue data, are available online, which enables the development of data-intensive applications. We developed the Continuous Dialogue Corpus Creation methodology and the corresponding technical infrastructure. The method enables not only new interoperable dialogue resources creation, but also available dialogue resources consolidation, and when annotated with dialogue act information to map and convert them into resources annotated with the standard semantic concepts defined within the ISO linguistic annotation framework. The corpus is used as a shared repository for analysis and modelling of interactive dialogue behaviour, and for implementation, integration and evaluation of dialogue system components. These activities are supported by the use of ISO standard data models including annotation schemes, encoding formats, tools and architectures. Standards facilitate practical work in dialogue system implementation, deployment, evaluation and re-training, and enable automatic generation of adequate system behaviour from the data. The proposed methodology is applied to the data-driven design of two multimodal interactive applications - the Virtual Negotiation Coach, used for the training of metacognitive skills in a multi-issue bargaining setting, and the Virtual Debate Coach, used for the training of debate skills in political contexts. Two interoperable dialogue corpora were constructed and released to the research community - the Multi-Issue Bargaining and Debate Trainee Corpora.

The third and main contribution is concerned with the incorporation of advanced cognitive models of adaptive, multitasking and human learning behaviour into a multimodal dialogue system, more specifically into dialogue management architectures. We designed a flexible and adaptable multimodal dialogue management system driven by cognitive modelling of human interactive, adaptive and learning behaviour. The presented approach to dialogue management integrates basic and advanced cognitive task agents able to reason about the behaviour, goals and strategies of human partners engaged in a debate or negotiation tasks. The implementation makes use of a theoretical novelty in Instance-Based Learning for Theory of Mind skills and integrating this in the dialogue management of cognitive tutoring system. The Debate and Negotiation Task Agents leverage established cognitive theories, namely Cognitive Task Analysis (CTA) and ACT-R simulations, to generate plausible, flexible behaviour in rather complex multimodal settings. The multi-agent Dialogue Manager proposes a flexible architecture separating modelling and processing of task-related and dialogue control actions which is beneficial for the current and future dialogue system designs. This work was successful: the dialogue system with the integrated cognitive agent technology delivers plausible task-related decision taking behaviour leading to reasonable user acceptance and satisfaction.

The dialogue system architecture is designed to be used in a mutiperspective setting: as an Observer, a Mirrorer, an Experiencer and as a Tutor. A modular architecture was designed. For this, a flexible technical integration framework for a modular distributed dialogue system is proposed. The developed architecture is open since it is not limited to one application domain, use case or technical solution: the core components are applicable outside the negotiation and/or debate domain (e.g. medical, human resource management, etc.), the proposed architecture can be extended to other use cases (e.g. customer support management training) and to novel processing algorithms and emerging HCI and AI technologies. Additionally, other modern devices and sensors (e.g. GPS positioning, web cameras, eye-trackers, biometric sensors, etc.) could be considered for future extensions.

Finally, a methodology for multimodal dialogue system evaluation is proposed in terms of measuring system usability. We related the relative contribution of various objective parameters and subjective factors to quantify the usability of a dialogue system as proposed by ISO 9241-11 and ISO/IEC 9126-4 standards. The proposed usability approach provides a useful decomposition of the usability concept into several factors enabling a clear mapping of system performance to distinctive usability perception aspects. Factors have been established experimentally by collecting human judgments and testing internal consistency of the selected dimensions. The approach has the advantage of assessing the impact of different items on usability perception instead of simply summing up or averaging to compute an overall satisfaction score.

The project results could be profitably used for dialogue management design tools as a component of user-interface design in multimodal applications. New information obtained at each stage of the project contributes to the development of new or improvement of existing annotation querying, conversion and corpus creation tools for multimodal dialogue resources; incorporated trained classifiers for automatic dialogue act recognition; dialogue manager with incorporated cognitive task agents; and a set of data collection, tracking and

management tools. The project specifically contributes to the development of the next generation of multimodal dialogue systems which incorporate metacognitive control in order to enable adaptive, re-, inter- and pro-active behaviour in setting goals, choosing appropriate strategies and monitoring processes across domains and contexts. This allows efficient interactions with human users by increasing the system's flexibility and knowledge richness, simulating human-like strategic decision making process when balancing between cooperation and competition. Finally, the project also contributes to the mainstream complex AI modelling methods, providing cognitive interactive agent technology for incorporation into dialogue systems and other interactive environments.

1.5 Thesis outline

The thesis is organised in the following way.

Chapter 2 is concerned with the theoretical and empirical aspects of dialogue modelling. Fundamental notions of dialogue theory are reviewed, the concept of dialogue act is introduced, important aspects of multifunctionality, multimodality and grounding are discussed. Exiting alternative approaches to computational dialogue modelling are described. The semantic framework of Dynamic Interpretation Theory is presented.

Chapter 3 discusses the cognitive modelling framework. We specify the cognitive modelling task, with a focus on human interactive multitasking, learning and adaptive behaviour. We present cognitive models for task analysis, decomposition, prediction and learning, based on Hierarchical Task Analysis and developed using the ACT-R cognitive architecture with details for its key components and functions used in our study. We provide details on the instance-based and reinforcement learning models related to their decisions-making support, and the ability to generate adaptive task-related behaviour.

Chapter 4 addresses the data-driven dialogue system design. The study presents the Continuous Dialogue Corpus Creation (D3C) methodology. We discuss the main principles and key processes related to the corpus development which serves as a shared repository for the data-driven system design. The methodology is based on existing standard data models, in particular on the ISO 26417-2 data model introducing the basic dialogue concepts and the Dialogue Act Markup Language (DiAML) as the main corpus annotation and exchange format between system components. The proposed approach is illustrated by applying it to recent corpus creation activities when designing two different applications - Virtual Debate and Negotiation Coaches, see also Chapter 6.

Chapter 5 presents an approach to flexible and adaptive dialogue management driven by cognitive modelling of human dialogue behaviour. We apply the Information State Update (ISU) machinery that operates on a multidimensional context model. This approach not only captures the behaviour of dialogue participants adequately, but also enables the generation of flexible multimodal behaviour by the system, addressing various task-specific and interactive goals and expectations simultaneously. Cognitive Task Agents are integrated into Multi-Agent Dialogue Management. One task agent is the baseline CTA-based agent that deploys the expert-based hierarchical task analysis method and features the basic

functionality necessary for the dialogue system to play the roles of an Observer, a Mirror or a Tutor. The other task agent deploys instance-based learning to decide about its own actions and to reflect on the behaviour of the opponent, and acts as a Negotiator, a Mirrorer and a Tutor simultaneously. We show that task-related actions can be handled by Cognitive Task Agents to play multiple roles including those of a plausible dialogue partner. Separating task-related and dialogue control actions enables the application of sophisticated models along with flexible architecture in which various (including alternative) modelling approaches can be combined. The approach leads to a knowledge-rich representation of the participants' information states and flexible dialogue management strategies. Moreover, it offers possibilities for various future extensions, as we illustrate by examples from the debating skills training scenario. The dialogue system together with human actors participated in a (meta)cognitive skills training within a debate and a negotiation based scenario.

Chapter 6 is concerned with the implementation and evaluation of two interactive tutoring applications - Virtual Negotiation and Virtual Debate Coaches. We discuss the system key components of the integrated dialogue system related to multimodal signal recognition, interpretation, management and generation. The technical integration framework for a distributed modular system is proposed. Inter-module communication design is detailed. The proof of concept systems are evaluated in trainee-based settings.

Chapter 7 draws conclusions from the main findings of the thesis, and sketches perspectives for future research on the basis of our results.

Chapter 2

Dialogue modelling

This chapter introduces the fundamental concepts important for computational dialogue modelling. These concepts are related to the basic dialogue phenomena and regularities observed in natural multimodal behaviour of dialogue participants. Dialogue participation in this work is treated as complex collaborative communicative activity. It is multifunctional, multimodal and multi-tasking. We discuss existing alternative approaches to computational dialogue modelling as a basis for human-computer dialogue system design. Finally, we present Dynamic Interpretation Theory (DIT) as the theoretical framework used in this thesis for its multidimensional view on dialogue and its formal definitions of dialogue acts in terms of update semantics.

Introduction

Dialogue models provide the basis for the interpretation of the participants' dialogue behaviour and for the decisions concerning the system's future actions. The design of a dialogue system that aims to exhibit rich multimodal interactive behaviour, starts in the first place with gaining a good understanding of human natural dialogue behaviour and adequate modelling of it. This involves answering a number of questions related to

- aspects of participating in dialogue: what are participants' tasks, roles and associated contributions to dialogue? what are these contributions motivated/triggered by?
- qualities of participating in dialogue: what is the right interpretation of these contributions? what are the factors that influence this interpretation?
- facets of participating in dialogue: what governs human dialogue behaviour? how are the governing principles related to observable linguistic, paralinguistic and extralinguistic features of such behaviour?
- aspects of involvement in dialogue: how is the dialogue structured? what are protocols and conventions that participants adopt?

The view taken in this work is that dialogue participation is a rather complex activity and should be modelled as such. Complex in the sense that people involved in dialogue do not only perform task-related actions. As many observations show, these actions are only a relatively small part of what happens in natural conversation (i.e. no more than 40% of all dialogue contributions are performed for this purpose, see e.g. Petukhova, 2011). Among other things, dialogue participants have constantly to evaluate whether and how they can (and/or wish to) continue, perceive, understand and react to each other's intentions. They share information about the processing of each other's messages, elicit feedback, take turns, monitor contact and attention, etc. People often have multiple goals in dialogue communication, they can communicate effectively and efficiently because they use linguistic and nonverbal elements in order to address several aspects of the communication at the same time. Complex also in the sense that people use several modalities when interacting with each other. Human interactions are more than the exchange of information, decision making or problem-solving; they involve a wide range of aspects related to feelings, emotions, social status, power, and interpersonal relations, and the context. Such dynamics is observed in human multimodal communicative behaviour, and adequate modelling of relevant multimodal dialogue aspects needs to be addressed. The computational dialogue model needs to be flexible and elaborate enough to deal with several complexities of human dialogue.

Challenges in natural human dialogue modelling addressed in this and the next chapters are related to frequent dialogue phenomena such as participants' multimodal dialogue actions, multi-tasking dialogue behaviour, multifunctionality of dialogue contributions, contents of participant's mental states and processes of their creation, grounding and learning. This and the next chapter serve as the theoretical background for analyses and developments presented later in this thesis, introducing fundamental concepts that play a key role in this study.

In this chapter, we first discuss the kinds of meaning that can be distinguished in dialogue, bringing us to the discussion of the notion of 'dialogue act' (Section 2.1). Section 2.2 addresses the phenomenon of multifunctionality of dialogue utterances and multi-tasking of human dialogue behaviour that motivates its parallel processing by the dialogue system. Section 2.3 discusses the multimodal aspects of natural human dialogue behaviour and their impact of participants' mental states which correspond not only to thinking but also to feelings, constituting mental representations and attitudes. In Section 2.4, we discuss humans' ability to use their background task knowledge and assumptions about their partners' mental attitudes and information states, and to share and 'ground' this knowledge for a successful and efficient dialogue. Section 2.5 introduces existing widely used approaches to dialogue modelling starting with dialogue grammars and ending with the most recent end-to-end neural network based models. Section 2.6 concludes the chapter by presenting the theoretical framework of Dynamic Interpretation Theory (DIT) used in this study. DIT offers a multidimensional dialogue model supporting an accurate understanding of multimodal multi-tasking adaptive behaviour which can be tuned to various dialogue situations. DIT provides precise formal definitions of dialogue acts as update operators on the dialogue context.

2.1 Dialogue acts

In order to understand and to describe what is happening in dialogue it has become common to analyse dialogue contributions in terms of communicative acts performed by a speaker. The idea of interpreting dialogue behaviour in terms of communicative actions such as greetings, statements, questions, promises and requests goes back to speech act theory (Austin, 1962; Searle, 1969). Before, in the line of logical positivism (see e.g. Ayer, 1966), it was assumed that meaning of a dialogue utterance can be captured by logical formulae that describe facts or 'state of affairs' that an utterance expresses and can be verified as true or false. However, as many researchers noticed, not all (if not the majority of) utterances are not truth-conditional statements but are a kind of actions. Speech act theory has been an important source of inspiration for modern dialogue act theory. The idea behind the theory of speech acts is to analyse natural language utterances in terms of actions performed by the speaker. According to Austin (1962), speech acts can be analysed at three levels:

1. a *locutionary* act, the performance of an utterance: the actual utterance and its meaning, comprising phonetic, phatic and rhetic acts corresponding to the verbal, syntactic and semantic aspects of any meaningful utterance;

2. an *illocutionary* act: the pragmatic 'illocutionary force' of the utterance, thus its intended significance as a communicative action;

3. and a *perlocutionary* act: its consequences and effects on the addressee, such as convincing, scaring, or getting someone to do or realise something.

While speech act theory is primarily an action-based approach to meaning within the philosophy of language, dialogue act theory is an empirically-based approach to the computational modeling of communication, in particular of linguistic and/or nonverbal communicative behaviour in dialogue. Dialogue acts are semantic concepts used to describe meaning of communicative behaviour. They can be defined by the way they are intended to affect the information state of an addressee when (s)he understands the behaviour.

Allwood (1977) notices that the identity of a communicative action should be determined in exactly the same way as the identity of any other action. He sees an action as combination of:

- *intention* and purpose that an agent connects with an action;
- *behavioral form* an agent exhibits in performing an action (e.g. linguistic form);
- *effects* or results of a certain type of behavior;
- *context*, because an action of a specific type occurs in a certain context.

In any communicative situation, interlocutors communicate their beliefs, desires, expectations, interests and obligations by means of certain communicative actions, i.e. dialogue acts. These actions are used by the speaker to signal his/her intentions concerning events,

objects, relations, properties involved in the communicative situation. For instance, when an addressee understands the utterance *Do you know what time it is?* as a question about the time, then the addressee's information state is updated to contain (among other things) the information that the speaker does not know what time it is and would like to know that. If, by contrast, an addressee understands that the speaker used the utterance to reproach the addressee for being late, then the addressee's information state is updated to include (among other things) the information that the speaker *does* know what time it is. Distinctions such as that between a question and a reproach concern the *communicative function* of a dialogue act; the objects, properties, relations, events, etc. that are referred to, constitute its *semantic content*. The communicative function of a dialogue act specifies how an addressee should update his/her information state with the information expressed in the semantic content, when (s)he understands the speaker's dialogue act correctly (i.e., as intended by the speaker).

In formal dialogue theories, actions are usually seen as transitions from state to state, with dialogue acts as special cases of actions. These theories define dialogue acts as having different sorts of effects on the dialogue context, mental states, or social context, see e.g. Bunt, 1994 and 2000; Poesio and Traum, 1998; Cooper, 2004. Several sets are associated with actions: (1) a set of effects or constraints on the resulting state, (2) a set of preconditions which are constraints on the initial state, and (3) sets of decompositions, i.e. sub-actions that performed together constitute the action (Traum, 2000). Effects correspond to achieved result(-s), contextual aspects and intention are related to the preconditions, and the form of behavior is characterised by the decompositions. There are three aspects of context as potential conditions that could be relevant for defining dialogue act types: dialogue state encoded in dialogue grammar (Traum and Allen, 1992; Lewin, 1998) or structural representation of context (Ginzburg, 1998); planning in terms of mental states of the speaker and addressee (*beliefs* and *intentions*, e.g. in Allen and Perrault, 1980); and the third one is defined in terms of the social obligations and commitments undertaken by the dialogue participants (Allwood, 1994). Most approaches combine two or three of these kinds of conditions and effects. For example, Dynamic Interpretation Theory (DIT; Bunt, 1994; 2000) models communicative agents as structures of goals, beliefs, preferences, expectations, and other types of information, plus memory and processing capabilities such as perception, reasoning, understanding, planning, etc. Part of these structures is dynamic in the sense of changing during a dialogue as the result of the agents perceiving and understanding each other's communicative behavior, of reasoning with the outcomes of these processes, and of planning communicative and other acts (Bunt, 2000). DIT is a context-change approach to dialogue acts and considers utterance meaning as defined in terms of how they affect the context. Context is used to refer to all factors that may be relevant for the understanding of communicative behavior. See Sections 2.4 and 2.6 for a general description and Section 5.2 for formalisation and implementation.

Thus, analysing the meaning of a dialogue utterance two fundamental aspects are distinguished: *semantic content* and *communicative function*. Informally speaking, a dialogue act is an act of communicative behaviour performed for some purpose. A formal interpretation of a dialogue act can be given when viewing the combination of a communicative function

and a semantic content as an operation that updates the information states of the dialogues participants in a certain way. A dialogue act is (Bunt, 1994):

(1) *a unit in the semantic description of communicative behavior in dialogues, specifying how the behavior is intended to change the information state of the addressee through his interpretation of the behavior.*

Dialogue acts are used to characterise communicative behavior in dialogue and should have an empirical basis. In other words, a dialogue act type should be reflected in observable features of communicative behavior. There are two criteria to distinguish a particular type of dialogue act (Bunt, 2000):

1. it corresponds to a specific context-changing effect;
2. the intended context-changing effect can be indicated by means of certain observable features of communicative behavior.

Dialogue acts are central in theories of dialogue and are often used in studies of dialogue phenomena, in describing the interpretation of communicative behaviour of participants in dialogue, and in the design of dialogue systems. Chapters 4, 5 and 6 address the latter two roles and usages of dialogue acts.

2.2 Multifunctionality, multitasking and parallel processing

In order for the system to enable smooth and robust processing and interpretation of all the information obtained from tracking and recognition devices, parallel processing of multiple hypotheses generated by different modules needs to be allowed by the dialogue system. Evidence suggests that understanding involves parallel generation of multiple hypotheses. In human processing, all possible hypotheses are activated in parallel until it is possible to identify a single candidate or reduce their number as has been shown, e.g. for processing ambiguous words by Swinney (1979) and Simpson (1994), for definite expression resolution (Tanenhaus et al., 1995), and for pronoun interpretation (Corbett and Chang, 1983).

Participation in dialogue is a complex and inherently multi-tasking activity. Every dialogue is motivated by a task, often a non-communicative one like solving a particular problem, plan some actions, etc. Dialogue may have a pure communicative underlying task like chat with your friends to maintain a relationship. While performing these task(-s), dialogue participants need to connect and organise ideas, fill gaps in their knowledge structures, evaluate evidence, argue with new information, test and modify, predict, clarify, generate questions, learn new concepts, make unexpected connections, reflect, analyse, synthesise and loop back. Additionally, to be successful dialogue participants need to perform multiple communicative and interactive tasks: ensuring contact, providing feedback, monitoring attention, taking and giving turns, repairing communicative failures, etc. At the core, a dialogue participant has three tasks: (1) to monitor own and partner's dialogue behaviour; (2) to understand partner dialogue contributions (i.e. intentions); and (3) to react adequately

to partner's intentions by performing suitable actions to pursue a certain task. Successful dialogue interaction involves understanding perceptions, cognitive, meta-cognitive and emotional processes by both partners.

As a result, dialogue participants often use linguistic and nonverbal elements in order to address several interactive and task-related aspects at the same time, and the majority of dialogue utterances are multifunctional, see Bunt (2007) and Petukhova (2011). For example[1]:

(2) C1: I would suggest we do not allow smoking in public places
B1: Uh-uhu
C2: What do you think?
B2: Uhm... yeah ... it's a bit difficult for me

Speaker ($C1$) produces a dialogue act with the communicative function of Suggest in the Task dimension, i.e. $\langle Task; suggest \rangle$. Speaker (B1) acknowledges that a dialogue act $C1$ is performed, and B1 may also have a function of stalling for time. We also can interpret B1 as understanding C1 as a suggestion and accepting it. Speaker C continues and perform a C2 dialogue act with the communicative function $\langle Task; setQuestion \rangle$ inviting the partner to react to C's suggestion: accept, reject it or propose a counter-suggestion. Utterance B2 is highly multifunctional. The small segment *Uhm...* is not really part of neither *yeah ...* nor the answer *it's a bit difficult for me*. It is a Turn Accepting and a Stalling act simultaneously. Similarly, *yeah ...* is multifunctional and expresses positive Auto-Feedback and Stalling acts. The segment *it's a bit difficult for me* is the answer to the question C2, but also in B2 suggestion C1 is declined. Thus, the utterance B2 in (2) is analysed as consisting of four functional segments: the overlapping segment one corresponding to an Answer to C2 and DeclineSuggestion of C1, the segment *Uhm...* corresponding to a Turn Accepting and a Time Stalling act, and the segment *yeah ...*corresponding to positive Auto-Feedback and a Time Stalling act. By DeclineSuggest expressed in B2, a tentative interpretation of B1 as an AcceptSuggest act is canceled.

Additionally, there might be multiple strategies to segment dialogue into meaningful dialogue units and this in multiple modalities. In natural conversation the use of speech is combined with nonverbal signs and vocal sounds, and all participants are most of the time performing some nonverbal communicative activity. DIT, and its subset ISO 24617-2 [ISO, 2012], allows multiple segmentation. Communicative functions can be assigned in multiple dimensions to units called functional segments, which are defined as the functionally relevant minimal stretches of communicative behaviour (see Geertzen et al., 2007). Figure 2.1 illustrates the segmentation and annotation of multimodal units.

Allwood (1992) has claimed that an utterance in dialogue tends to be both sequentially and simultaneously multifunctional. Bunt (2007) proved empirically that whatever segmentation method is used and whatever annotation strategy, multifunctionality never goes away. When a coarse segmentation method is used, that considers entire turns as the units of communicative behaviour, it was found that the functional units have on average five communicative functions, about half of which is due to sequential multifunctionality and half

[1] From Metalogue Multi-issue Bargaining Corpus, see Petukhova et al., 2016

Speaker		Observed communicative behaviour											
A	*words*	Uhm	I will be talking about			Two funda-		uh		Three fundamental			ideas
	gaze	averted	personB				averted			personB			averted
	head									single nod			
	hand					2 fingers up				3 fingers up			
	posture	working position				random shifts (dancing)							
	Discourse Structuring		Inform							Inform			
	TurnM.	Turn-				Turn-keep							
	OCM									Self-Correct			
B	*words*												Okay
	gaze	personA	averted			personA							averted
	head							Sideway single movement				Single short nod	
	eyes				narrow								blinking
	lips							Random movements					
	Discourse Structuring												Agreement
	Auto-FB				Neg. execution							Pos. understanding	
	TurnM							Turn-grab				Turn-take	

Figure 2.1: Transcription, multimodal segmentation and annotation across multiple dimensions. From the Metalogue Debate Trainee Corpus (Petukhova et al., 2018).

to the multifunctionality of the smallest possible functional units that may be distinguished within a turn. These smallest possible units, which correspond to the functional segments of a multidimensional segmentation, have usually two or three communicative functions; functions for Turn Management are responsible for about half of this.

If we consider multimodal segments, the previously performed analysis showed that nonverbal communicative behaviour may contribute to the multifunctionality of dialogue utterances by (1) emphasising or supporting the communicative functions of synchronous verbal behaviour (i.e. qualifying it); (2) performing separate dialogue acts in parallel to what is contributed by the partner; and (3) expressing a separate communicative function in parallel to what the same speaker is expressing verbally (Petukhova, 2010).

The interpretation and generation hypothesis space is potentially big when computing information from all parallel processing modules concerned with Automatic Speech Recognition (ASR), Natural Language Understanding (NLU) dealing with lexical analysis (lexical lookup, possibly supported by morphological processing, and by additional resources such as WordNet, VerbNet, or lexical ontologies), syntactic and semantic parsing (construction of syntactic interpretations, computation of propositional, referential, or action-related content), Visual Motion Interpretation (VMI) and Dialogue Act Recognition (DAR) when determining speaker implicit and explicit intentions, and considering multiple segments. To deal with this may require additional memory, strategies and methods.

The following scenarios for behaviour interpretation are possible and some of them are simulated by our system when performing the dialogue act recognition task, see Section 6.1 for details:

- (a) generation of multiple interpretations in parallel, all but one get eliminated at the end;
- (b) generation of multiple interpretations in parallel, keep all active;

- (c) generation of one hypothesis that is constantly refined; and
- (d) no hypothesis generation but postpone the decision till the end of the relevant segment.

2.3 Multimodality, affected states and social signals

Human natural and modern human-computer interactive technology are multimodal, driven by the modes involving the five human senses: sight (vision), hearing (audition), taste (gustation), smell (olfaction), and touch (somatosensation). Modalities that are commonly used include speech, gestures (both on-screen "touch" gestures and in-air gestures), facial expressions, eye gaze and haptics. The current state of technology enables tracking of body movements, head, hand and arm gestures, gaze direction and facial expressions, i.e. most laptop computers, tablets, and mobile phones are already equipped with cameras, microphones and speakers as well as with rather sophisticated sensing devices. Many operating systems incorporate speech processing facilities (Siri, GoogleSpeech, MicrosoftTellMe, Cortana, etc.). The current state of technology enables fine grained and inexpensive tracking of visible body movement and facial expressions (Intel®RealSense™, 3D Kinect; eye-trackers like Tobii Glasses) and various biometrical signals (Blood Volume Pulse and NeXus EXG sensors). Social robots are situated and human-like. Novel devices and sensing technologies get more and more interconnected and seamlessly integrated in everyday human activities. Multimodal conversational interfaces enable users to interact with their devices, appliances and other systems in an intuitive and natural manner. Multimodal dialogue is not only the most social and natural form of interaction, but has been proven to have positive effects when incorporated in human learning and medical treatment, see e.g. Sali et al., 2010; Woods et al., 2012; Hughes et al., 2013. It has also been shown that 'digital immersion' can enhance learning and increase user acceptance (Dede, 2009; Sadowski and Stanney, 2002; Lessiter et al., 2001). Multi-sensory approaches not only reinforce learning, but also personalise the assessment process and engage learners.

While exhaustive real-time monitoring seems unrealistic, certain multimodal markers may be defined that trigger and guide the interaction and presentation of information. Progress has been made in multimodal behaviour modelling, with advances in social signal processing and affective computing, see Vinciarelli et al., 2009 for an overview. It seems to be plausible to define multimodal (including biometric) markers that can guide the presentation of information and interaction in a principled way, and then seek these particular markers as triggered by the presented material. To give a concrete example, facial signs for boredom can be analysed at places where a particular unit of presentation is longer than average. Similarly, uncertainty can be detected from facial expressions, but also from analysing user's typing behaviour (e.g. typing speed, numbers of deletions, substitutions, insertions and other corrections) or from the interaction with interfaces (e.g. mouse movement patterns, clicks, backtracking actions, etc., see Van Dam, 2006). If boredom or uncertainty is detected, the system may take certain intervening actions through comments, recommendations, extra content, etc.; else, it continues presenting the material as originally planned. Thus, the

way in which the user is performing his tasks and reacts to system actions will not only be linked to his knowledge, skills and task performance, but also to his engagement and motivation. Based on the motivational assessment appropriate system interventions are triggered, if suitable, to support and keep up users' engagement e.g. encouraging feedback or attention catchers (Steiner et al., 2009).

Aside from motivation, emotions play a key role in many if not all interactive contexts. Several psycho-pedagogical theoretical approaches focus on the psychological aspects and feedback loops between emotion, motivation, interaction and learning. For example, the Control-Value Theory of Achievement Emotions (Pekrun and Perry, 2014) proposes a framework for the different antecedents and effects of achievement emotions and their interrelations with motivation and learning. Thereby, emotions are defined as a multidimensional construct with affective, cognitive, motivational, expressive and peripheral physiological processes (Damasio, 2004). For an analysis of a user's affective state it is crucial to have appropriate measurements and indicators. According to a compositional model of emotion that conceptualises emotions by experiential, physiological, and behavioural components, there are various possibilities of measuring emotions (Mauss and Robinson, 2009). Traditional questionnaires are only of limited value since they are normally retrospective and disrupt the ongoing interactive process. Even though some scales exist that allow a very short and immediate assessment of emotions (like the Smileyometer of Jäger, 2004), they nevertheless disrupt the user. Accordingly, other assessment methods are necessary. One possibility is to exploit information on user behaviour via log files for affected state analysis. They can be combined with questionnaire data (Linek et al., 2008).

There is a wealth of research performed on the psycho-physiological measurement of emotions and automatic emotion recognition from speech and visual signals. Eye-tracking data, for example, delivers not only information about the users' attention by means of frequency and duration of gaze fixation on the Areas of Interest (AoI), but also provides evidence about the positive versus negative emotional reaction on the fixated object via the pupil size. Based on the general assumptions of the well-established Facial Acting Coding System (FACS; Ekman and Friesen, 1978; Cohn and Ekman, 2005; FACES by Kring and Sloan, 2007) facial muscle contraction and the related indicators of facial electromyography (EMG) may serve as a source of information for the analysis of the user's affected state (Tassinary and Cacioppo, 2000). It is broadly accepted that a single psycho-physiological indicator is insufficient for the assessment of a specific emotion. Rather, a pattern of different indicators and probably additional indicators should be used (Barrett et al., 2007; Kreibig et al., 2007; Larsen and Prizmic-Larsen, 2006; Mauss and Robinson, 2009).

Social signal processing is a new but rapidly maturing branch of computer science which aims at an understanding and modelling of human social interactions for providing computers with similar abilities for use in human-computer interaction scenarios. A social signal is a communicative or informative signal that, either directly or indirectly, provides information about social facts concerning interactions, emotions, attitudes, or social relations.

The processing of social signals in video or audio material enables identification and conceptualisation of social signalling patterns that are stable at least for a given context and

culture. It enables detection and understanding of nonverbal behavioural cues conveying social signals. These signals are implicit in video or audio recordings, and allows us to synthesise similar nonverbal behavioural cues conveying desired social signals for embodiment of social behaviours in output representations for summarising speaker intentions and cognitive states.

Computer systems and devices capable of sensing agreement, inattention, or dispute, and capable of adapting and responding in real-time to these social signals in a polite, non-intrusive, or persuasive manner, are likely to be perceived as more natural, efficacious and trustworthy. Human interactions are more than the exchange of information and offers, decision making or problem-solving; they involve a wide range of aspects related to feelings, emotions, social status, power, and interpersonal relations, and the context. For many real-life interactive situations, it is important for people to maintain good relations, build trust over time. In contrast, social barriers can trigger interactive processes that lead to bad communication, polarisation and conflict escalation. Successful interlocutors acknowledge social signals and react to them. Digital conversational agents need to do the same by employing tools that can accurately sense and interpret social signals and social contexts, learn context-dependent social behaviour, and use it properly (e.g., see Pelachaud et al, 2002). The research results in the field attest that social interactions and behaviours, although complex and rooted in the deepest aspects of human psychology, can be analysed automatically with the help of computers.

Most modalities are symmetric in the sense that they can be used for input as well as for output (although possibly via different hardware). For generation, for example, combining synthetic speech with laughter influences the perception of social bonds (Trouvain and Schröder, 2004). Similarly, facial expressions influence a human user's evaluation of an Embodied Conversational Agent (ECA, see Ruttkay and Pelachaud, 2004). Signal contingency plays a key role in creating rapport between human user and virtual agent (Gratch et al., 2007). Politeness cues (Wang et al., 2005) and empathic expressions (Niewiadomski et al., 2008) are perceived as more appropriate in many interactive scenarios.

Generally, it is possible for a dialogue management module to remain largely "agnostic" with respect to input and output modalities. However, there is a need for a component that takes on the task of abstracting away the semantic content from the modality and to do multimodal fusion.

From a human science standpoint, language is the social signal par excellence, and should obviously be included. Technologically, there is an obvious motive to avoid it. Research findings, e.g. reported by Ambady and Rosenthal (1992), indicate that linguistic messages are rather unreliable means to analyse human behaviour, and it is very difficult to anticipate a speaker-dependent word choice and the associated intent in affective and socially-situated expressions. In addition, the association between linguistic content and behaviour (e.g. emotion) is language-dependent and generalising from one language to another is very difficult to achieve. By including this information in a dialogue model we expect to provide the necessary link.

2.4 Dialogue context and grounding

Humans are able to make sense of a dialogue even when little linguistic information is present in the partner's dialogue utterances. They act as cognitive agents with the abilities to perform inferences based on available background knowledge and experiences, and assumptions on the partners' mental states. Mental states are hypothetical states that correspond to thinking and feeling, and consisting of speaker mental representations and propositional attitudes. Mental representations correspond to our mental image of the world and enable representing things that have never been experienced as well as things that do not exist. Propositional attitudes are relational mental states connecting a person to a proposition, and are often assumed to be the fundamental units of thought and their contents. In most computational dialogue modelling approaches (Isard, 1975; Bunt, 1989; Ginzburg, 1996), mental states are limited to the participants' intentions, desires, beliefs, expectations, etc. The mental state has also been conceived as a dialogue plan which includes goals, actions to be achieved and constraints on the plan execution, see e.g. Traum (1993). Some researchers consider mental states as equivalent to emotional states (Nisimura et al., 2006), given that affect is an evolutionary mechanism that plays a fundamental role in human interaction to adapt to the environment and carry out meaningful decision making (Callejas et al., 2011; Sobol-Shikler, 2011).

We consider a participants' information state[2] as the totality of conditions that influence the understanding and generation of dialogue interactive behaviour as defined by Bunt (1994). The structure of a participant's information state is potentially complex and consists of (but is possibly not limited to) linguistic, semantic, cognitive, physical and social contexts (parts).

To be successful in communication, participants should be able to integrate the meaning of dialogue contributions with the representation of the existing/available context, parts of which are dynamic, i.e. changing as the dialogue proceeds, other parts corresponding to global contextual properties of a dialogue setting and participant roles, and previously available shared knowledge about other dialogue participants. The integration process is collaborative, where participants coordinate their actions at many levels. The coordination of the beliefs and assumptions of the participants is a central issue in any communication. In other words, in the speaker role a participant produces his contributions in such a way that it can be correctly interpreted by the partners. To enable this, the speaker continuously monitors and evaluates whether addressee(-s) attend to, perceive, understand, and react to the speaker's intentions. The addressee gives his best to understand the speaker's utterances, react to their intentions, and report on his processing. Mutual or shared beliefs about understanding, created beliefs, knowledge, assumptions, presuppositions etc., are added to the participants' *common ground*, which is a set of propositions that the dialogue parti-

[2]In the literature, the tems 'information state' and 'mental state' are often used as synonyms. We use these two terms interchangeably as well, however, define the former formally considering it as a computational notion, while the later is mostly used as a more broader general concept. Note also that in this study as well as in many ISU-based approaches, the terms 'context' and 'context model' are considered as synonyms of 'information state'.

cipants mutually believe. The process of establishing and updating the common ground is called *grounding*. The grounding process involves the mentioning of facts and proposals in presence of other participants and then monitoring their understanding and update in the diagnosis step. If the information is processed successfully, this is signalled by positive feedback, otherwise processing failures are reported at different levels or addressees express their needs for additional explanations or clarifications.

While 'common ground' is not directly observable, grounding mechanisms are accessible through observable dialogue behavior, e.g. evidence of understanding what is said in dialogue is provided by feedback acts. The nature of such evidence depends on the communicative situation. In face-to-face conversation, for example, participants may present evidence of grounding through body movements and gaze re-direction, while in telephone conversations only verbal and vocal signals are available for the participants. It has been observed that not every participants' action is explicitly grounded or checked to be mutually understood, rather dialogue participants deploy the principle of *least collaborative effort* formulated by Clark and Wilkes-Gibbs (1986) as follows:

> In conversation the participants try to minimise their collaborative effort - the work that both do from the initiation of each contribution to its mutual acceptance.

Thus, if no (explicit) negative evidence arrives, the speaker assumes that the addressee's processing of his contributions was successful. The addressee may continue listening rather than letting the speaker know at each point in time how the addressee understands the speaker. The contributor and the partners mutually believe that the partners have understood what the contributor meant to a criterion sufficient for the current purpose (Clark and Schaefer, 1989). According to the Contribution Model (Clark and Schaefer, 1989), participants in dialogue perform collective actions ('contributions') that result in grounding. To make a contribution requires (1) content specification (a speaker tries to specify the content of his contribution, and the partners try to register that content), and (2) grounding (participants attempt to establish the mutual belief that they understand what was said). Each contribution has two phases: a *presentation* phase, where the speaker presents an utterance for the addressee to consider, and an *acceptance* phase, where the addressee gives evidence that he believes he understands what the speaker means by this utterance. Grounding may occur if the dialogue partners mutually believe that everyone involved has a clear enough understanding of what was said and accept it to move the dialogue forward. There are several ways by which dialogue participants may signal this: (1) by explicitly signalling acceptance, either verbally or by smiling, nodding, etc.; (2) by requested clarification and resolving misunderstandings; and (3) by moving forward with a new topic (and waiting to see if the partner expresses confusion).

The Contribution Model, however, does not specify what mutual beliefs are created and when, and how they are updated, see Traum (1994). The computational model of grounding defined as a function in advancing the mutual understanding is essential to design an adequate computational dialogue model. We base our developments on the work of Bunt et al.

(2007) who proposed their semantically motivated grounding model using the framework of Dynamic Interpretation Theory (DIT) (Bunt, 2000). Dialogue context (discussed above) and dialogue acts (see Section 2.1) are the main ingredients of this model. Information is transferred from one dialogue participant to another through belief creation (understanding) and belief transfer (adoption). The speaker expects under 'normal input-output' conditions (Searle, 1969) that what he is saying is perceived and understood as intended. These expectations may be strengthened when there is positive evidence from the audience, and if negative feedback arrives these expectations are canceled. The DIT grounding model and mechanisms will be discussed in more detail in Section 5.6 and illustrated by examples for our applications.

Optimal grounding strategies depend on the specifics of the communicative setting and the style of communication that is appropriate for that setting. For instance, in the case of dialogues involving the transfer of important information such as credit card numbers, it is desirable to give extensive and explicit feedback and do so in a uniform manner, but in more informal dialogues, persistent positive auto-feedback should rather be avoided. The choice in favour of one strategy over another also depends on the global communicative setting, e.g. possible negative physical, psychological or welfare consequences for a communicative partner (high risk vs low risk), available communicative channels and their quality (telephone vs face-to-face interaction) and so on. Grounding costs and strategies may also vary on features like

- Co-presence: participants are near each other, and can point at objects in common ground
- Co-temporality: participants can expect to receive a timely reply; interruptions or delays are significant
- Simultaneity: participants can send and receive at the same time; allow interruption, backchannel feedback
- Sequentiality: participants contributions are strictly ordered, and cannot get out of order
- Reviewability: participants can look at the past history of the conversation
- Revisability: participants have the option of editing their contributions before they commit to them

Given the state-of-art in speech recognition and natural language processing, spoken natural language based dialogue systems have much greater need for feedback, clarifications and corrections than appear in human-human interaction. On the other hand, including explicit feedback and verification after every user turn can make dialogue dull and inefficient. Optimising the amount of explicit positive, and the form and timing of negative feedback is an important challenge for a dialogue system. Our small-scale experiments showed that users generally appreciate diversity in system behaviour, ranking a system that

displays variations in behaviour higher than a system that follows simple sequential patterns or overgenerate explicit feedback in every system turn [Petukhova et al., 2015a]. Important aspects are addressed in Section 5.5.

2.5 Approaches to dialogue modelling

There are several paradigms for dialogue modelling and action planning for domains of varying complexity. The dialogue applications demand solutions to many problems of heterogeneous types, often having many common features. Three prominent approaches to dialogue modelling are *dialogue grammars*, *plan-based* approaches, and the *information state update* paradigm. Recently, data-driven statistical dialogue modelling, e.g. Partially Observable Markov Decision Process (POMDP, [Williams and Young, 2007]) based models, has gained a lot of attention in the community and has been successfully applied when designing spoken task-oriented dialogue systems.

2.5.1 Dialogue Grammars

Many traditional dialogue systems have been largely hand-coded proposing inflexible representations of dialogue states, implemented as some form of rule-based machine. For instance, the dialogue grammar approach is based on the observation that a dialogue exhibits certain regularities in terms of frequently occurring sequences of speech acts. For instance, questions are frequently followed by answers; requests and offers by acceptances or denials (Schegloff, 1968). Such *adjacency pairs* have been proposed to define grammar rules describing well-formed dialogues. So-called insertion sentences also frequently occur intervening between first and second part of the pair. An insertion sentence is topically related to the pair it interrupts and is mostly used to sort out/clarify information necessary to provide the second part. An example of dialogue grammar rules is provided in Figure 2.2.

Examples of dialogue systems that use a dialogue grammar are SUNDIAL (Andry et al., 1990; Bilange, 1991), LINLIN (Dahlbaeck and Jonsson, 1998) and RailTel (Bennacef et al., 1996). The dialogue grammar approach has been criticised for being far from providing adequate explanation of human natural dialogue behaviour. The model completely ignores (a) the semantic content of dialogue acts, and (b) the multifunctionality of dialogue utterances.

2.5.2 Finite-State Automata

In the Finite State Automaton (FSA) approach each action of a participant in a dialogue leads to a new state. Finite automata (transducers or Mealy automata) are simple algebraic structures that relate internal states to input and output sequences, offering a general model of the participation in a dialogue. A FSA-based protocol is a quintuple $\langle Q, q_0, F, \Sigma, \sigma \rangle$ consisting of a finite set of *dialogue states* Q including an *initial state* $q_0 \in Q$ and a set of *final states* $F \subseteq Q$, a *communication language* Σ, and a *transition function* $\sigma : Qx\Sigma \rightarrow Q$. The user is taken through the dialogue via following a sequence of pre-determined states. Consider the example of one simple FSA diagram in Figure 2.3 for a dialogue between two

Rule 1: There are no ongoing pairs. The system starts a new pair.

Rule 2: There is at least one ongoing pair. The user provides a response of the same type and on the same topic, thus completing the pair.

Rule 3: There is a single ongoing pair. The system provides a response of the same type and on the same topic. Then the system initiates a new pair of a possibly different type and on a possibly different topic.

Rule 4: There are at least two ongoing pairs on the same topic. So the dialogue must have entered an insertion sequence. The system provides a response to complete the most recent pair. The system reminds the user of the ongoing pair. The grammar achieves this by requiring that the system initiate a new pair of the same type and topic as the ongoing one but it does not push it onto the stack of ongoing pairs, which remains unchanged.

Rule 5: There is at least one ongoing pair. The user provides a response to complete the pair and initiates a new pair. This aborts any other ongoing pairs so the stack contains only the new pair.

Rule 6: There is at least one ongoing pair. The user aborts it and initiates something new. We know this not an insertion sequence because the topic is different.

Rule 7: There is at least one ongoing pair. The user begins an insertion sequence by not responding to the ongoing pair but by initiating a new pair on the same topic. Both pairs are now on the stack.

Figure 2.2: Dialogue Grammar rules of the Sermo recommender system (Bridge, 2002).

agents A and B, where A continuously informs B about c. An inform move uttered by agent A will take B from state 0 to state 1. Immediately after A has informed B, the latter can either choose to acknowledge that fact or he may end the dialogue. However, it would be illegal for A to continue the dialogue with another inform move unless he has received an acknowledgment from B first, and so forth.

In FSA, on one hand, the users' input is limited to pre-defined words or phrases, which may simplify speech recognition and natural language processing. Dialogue interactions are structured by directing the user through a dialogue, which also simplifies dialogue management and may result in rather reliable performance. FSA-based systems are straightforward to develop and are particularly suitable for well-structured tasks, e.g. hotel booking. Thus,

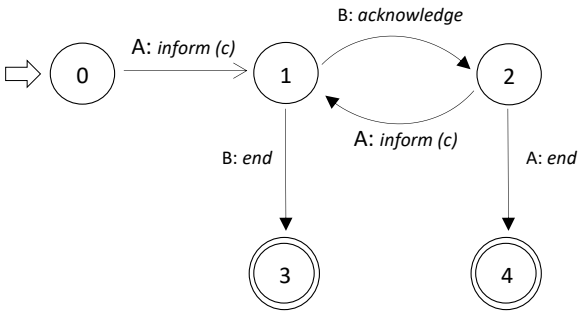

Figure 2.3: Diagram of a simple Finite State Automation.

the main advantage of dialogue scripting with FSA is the control over the dialogue progress and the determinism of the resulting system. This has advantages for dealing with the (spoken) input and output at any state. On the other hand, the users' limited input makes the system unable to handle more complex dialogues that deviate from simple pre-defined sequential structures. The approach requires explicit specification of error recovery and repair sub-dialogues, i.e. grounding behaviour must be hard-coded. The FSA models are inflexible, they do not allow users to take initiative and are not able to cope with unpredictable/unexpected user behaviour. Performing quite well on many restricted domains, some dialogue types, e.g. negotiations, cannot be modelled using FSA since participants goals and constraints often are not known in advance.

2.5.3 Frame-based approaches

A more realistic system than the finite state system is the frame based system or slot filling system. In this method, the system asks questions and answers from users are used to fill slots in the frame. This framework is inspired by frame semantics (Fillmore, 1977), where the meaning of natural language utterances are relativised to scenes. Frame semantics appears to provide a realistic and useful degree of abstraction. Any description of word meanings must begin by identifying such underlying conceptual structures. Frames have many properties of stereotyped scenarios – situations in which speakers expect certain events to occur and states to obtain. In general, frames encode a certain amount of "real-world knowledge" in schematised form. A frame is a comprehensive collection of concepts linked to each other.

Applied to dialogue modelling, the entire dialogue may constitute one frame or scene. Types of entities and their possible attributes involved in the scene are specified in a template. The task is conceived as filling a set of slots in a template, where the order in which slots are filled is not fixed beforehand. This allows for some degree of user initiative. The flow is not pre-determined but depends on the content of a user's input and the information the system has to elicit. A frame-based dialogue system typically asks questions of the user, filling the slots gleaned from user responses until it has enough information to perform a query. In this system, the user response can contain answers to multiple questions or slots and it is the duty of the dialogue manager to extract the necessary information from the user response and fill out the necessary slots while remembering not to ask questions for slots already filled out.

Entities in the application domain can be also hierarchically modelled where frames are arranged hierarchically to reflect the dependence of certain topics on others. For example, in Veldhuijzen van Zanten (1996), in the train timetable inquiry system OVIS[3], a frame structure relates the entities in the domain to one another, and this structure captures the meaning of all possible queries the user can make, see Figure 2.4.

Frame-based dialogue systems exhibit more flexible behaviour when comparing to FSA-based dialogue management. They allow for some degree of user initiative, concerning the

[3]OVIS is a Dutch acronym for Public Transport Information System, based on a system designed at Philips GmbH Forschungslaboratorien Aachen (Aust et al., 1994).

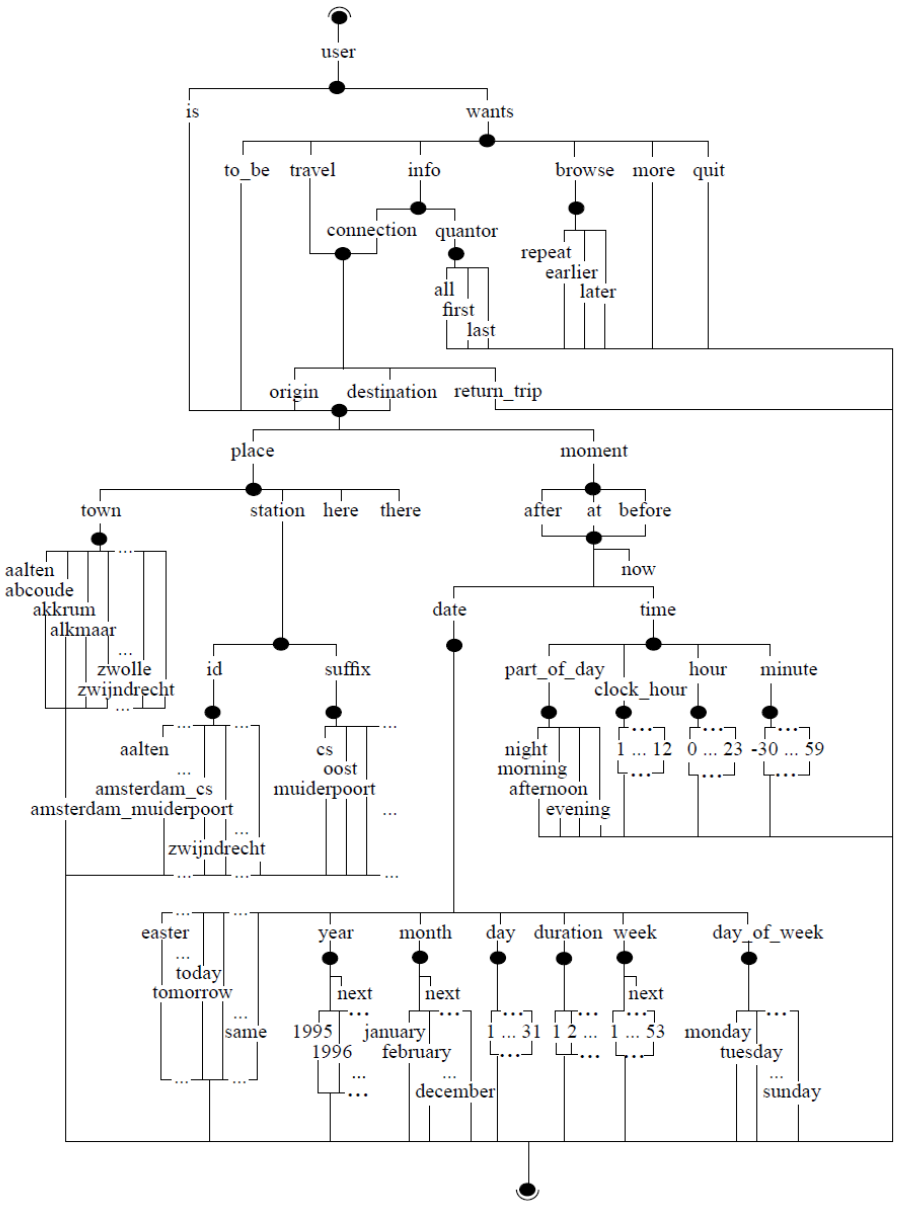

Figure 2.4: The frame structure for the OVIS system defined in Aust et al., 1994

order and the amount of information (slot-fillers) they provide in a single turn. However, it is not always possible to express the meaning of user utterances in terms of a rigid pre-defined frame structure.

2.5.4 Plan-based models

One of the most important tasks in dialogue management is deciding what to do/say next, i.e. selecting the next dialogue action of the system. One of the main factors that play a role deciding on the next system action is the task that is being addressed in the dialogue. Plan-based approaches to dialogue modelling are founded on the observation that participants in dialogue plan their actions to achieve certain goals. Dialogue participants have certain (sub)goals to accomplish and a progress has been made towards these (sub)goals with considerations what is the best way to carry on. The final global result forms a global intention, i.e. a *plan*. A plan is a course of actions which is intended to change the state of the world in a clearly identified manner. The generation of acts is determined by the plan and by the properties of the state of the world (context) in which actions will be executed. Thus, a plan is not a static but a dynamic set of actions which can be revisited during their execution.

Allen (1983) argues that people are rational agents, forming and executing plans to achieve their goals, and inferring the plans of other agents from observing their actions.

> It turns out that, if you want to coordinate and communicate with other agents, it is extremely useful and possibly even essential - for you and those other agents to be planners. There are two reasons for this. First, coordination between agents seems possible only because they can count on one another behaving in more or less stable ways, such as would result from an agent's commitment to its plans. . . . Second, . . . , communication is greatly facilitated by the agents reasoning about one another's plans. (Pollack, 1992)

The agent who has as a goal to carry out a certain plan may have several alternatives. Agents are capable to make choices among alternatives, i.e. take decisions or commit themselves to one of the possible alternatives, or if none is available or considered not possible for whatever reasons, to report this. Agents when generating plans aim at achieving a set of goals defined computationally in terms of directed search through a possible problem space with three main components: (1) a description of the current situation - *preconditions*; (2) a set of goals that the agent aims to achieve - *effects*; and (3) a set of actions by which the effects are achieved - *body*. Consider an example for an agent who has a plan to move a block in the Blocks World of the STRIPS (Stanford Research Institute Problem Solver) system [Nilsson and Fikes, 1970]:

(3) Name: move_block(Block, From, To)
Preconditions: on(Block, From), clear(Block_name), clear(To)
Effects:
Add list: on(Block, To), clear(From)
Deletion list: on (Block, From), clear(To).

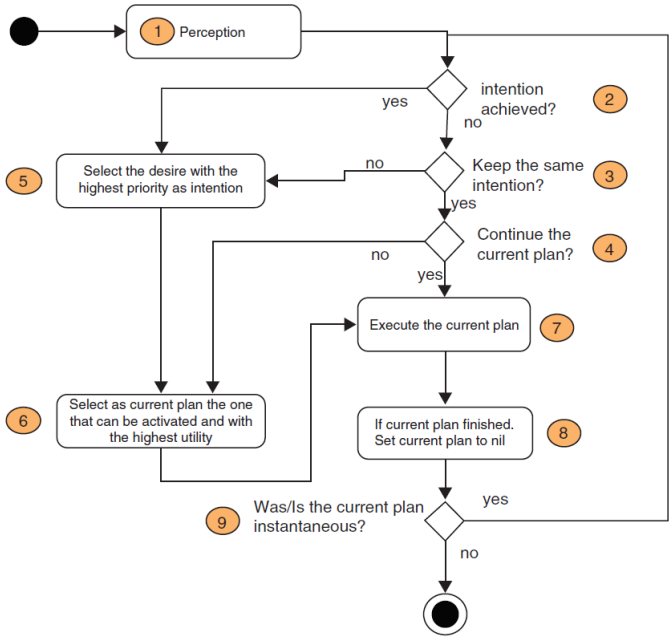

Figure 2.5: Activity diagram. Adopted from Caillou et al., 2015.

At each step, the agent will perceive the environment, then continue its current plan if it is not finished, or if the plan is finished and its current intention is not fulfilled, it selects a plan, or if its current intention is fulfilled, it selects a new desire to add to its intention stack. The applied 'thinking' process is schematically depicted in Figure 2.5.

2.5.5 BDI agent models

Complex tasks may require more sophistication, and ultimately, task planning requires problem-solving capabilities. Approaches to dialogue modelling employing Artificial Intelligence planning techniques are often referred to as *belief-desire-intention* (BDI) models, see e.g. Cohen and Perrault (1979), Allen and Perrault (1980), Sidner and Israel (1981), Carberry (1990) and Sadek (1991). A conversational agent components include perception, beliefs, desires, planning/reasoning, commitment, intentions and acting. The BDI model defines a computational architecture of rational agents represented in Figure 2.6.

Basically, two types of structure are defined that a participant's mental state contains: *beliefs*, consisting of an agent and a proposition which is believed by the agent, and *desires*, which represent the agent's goals. Belief is a type of mental (propositional) attitude which play a fundamental role in the BDI architecture. Beliefs are the main components of the agent's mental state and they are propositional attitudes held by an agent to be true. A desire can be intuitively defined as a state of the world the agent finds pleasant and that does not currently holds. Different desires may conflict with each others. *Intentions* are derived from desires since desire represent motivation for acting intentionally. Intentions arise from rational deliberation and they are future-directed: they reflect decision an agent

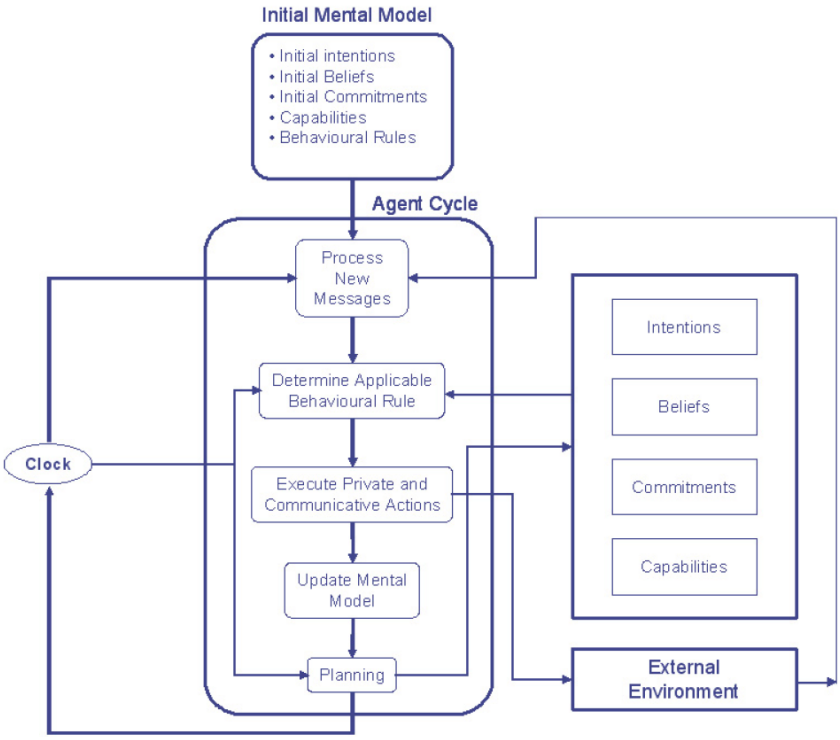

Figure 2.6: The BDI agent model. Adapted from Allen and Perrault, 1980.

	Request(Speaker,Hearer,Act)
CanDO.Pr	Hearer CanDo Act
Want.Pr	Speaker believe Speaker want$_{request-instance}$
Effect	Hearer believe Speaker want Act

Figure 2.7: Cohen and Perrault's definition of REQUEST.

has made about his future action. Intentions are modeled as plan operators. Figure 2.7 gives an example of how a Request is defined in terms of these operators.

In order to understand what the speaker is saying an addressee uses both utterance properties and clues from his model of the speaker's cognitive state in order to recognise the plan that made the speaker say what he said. Plan-based approaches relate a domain-level plan (e.g. a plan to get certain information, or to catch the train) with a communicative plan. Plan-based models assume a particular information flow for making inferences. First, a speech act is computed with its associated goal, then this information is used together with a domain plan to further specify the domain plan. A relationship between the current goal and the previous goal is constructed in order to infer implicatures of the current utterance, and therefore more information about the domain-level and communicative plans. This is what plan-based approaches are often criticised for. Plan construction and inference are activities that can easily get very complex and become computationally intractable. Moreover, some dialogue phenomena like actions that are not about planning or about the task at all, e.g. actions such as feedback, clarification questions, confirmations, etc., which constitute a great portion of all utterances in dialogue are difficult to model by means of plan recognition and plan generation. In order to overcome these shortcomings Grosz and Sidner (1986) and Grosz and Sidner (1990) proposed to consider conversation as a joint activity. According to this approach (known in literature as the *collaborative approach*) all dialogue partners work together to achieve and maintain understanding in dialogue. Collaborative approaches try to capture the motivation behind a dialogue and the mechanisms of dialogue itself, rather than focus on the structure of the task. This suggests that the beliefs of all dialogue parties should be modelled and if the proposed goal is accepted by another partner it will become part of the shared (mutual) beliefs (see also Traum, 1994 and Traum, 1999).

Plan-based models have been applied for example in the TRAINS system (Allen et al., 1994) and in the TRIPS system (Allen et al. (2001), which has a task manager that relies on planning and plan recognition. ViewGen (Wilks and Balim, 1991) is a system for modelling agents, their beliefs and their goals as part of a dialogue system, which uses a planner to simulate other agents' plans. Nested beliefs (about beliefs and goals) are created only when required as the plan is generated and are not pre-stored in advance before the plan is constructed, as in (Cohen and Perrault, 1979) and (Allen and Perrault, 1980). The Verbmobil speech-to-speech translation system uses a plan recogniser similar to that of plan-based models (Wahlster, 2000).

More recent work on plan-based dialogue modelling of Chu-Carroll and Carberry (2000) models dialogue with plans at several levels. While the type of levels differs with each model, all models include at least a domain level and a problem-solving level. In these

$$\begin{bmatrix} PRIVATE: & \begin{bmatrix} PLAN: StackSet(Action) \\ AGENDA: Stack(Action) \\ BEL: Set(Prop) \\ TMP: & \begin{bmatrix} BEL: Set(Prop) \\ QUD: Stack(Questions) \\ LU: & \begin{bmatrix} SPEAKER: PARTICIPANT \\ MOVES: assocSet(Move, Bool) \end{bmatrix} \end{bmatrix} \end{bmatrix} \\ SHARED: & \begin{bmatrix} BEL: Set(Prop) \\ QUD: Stack(Questions) \\ LU: & \begin{bmatrix} SPEAKER: PARTICIPANT \\ MOVES: assocSet(Move, Bool) \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

Figure 2.8: Example of information state as defined in Traum et al. (1999).

models, user utterances are interpreted as corresponding to certain actions within recipes at the various levels.

The major accomplishment of plan-based theories of dialogue is that they offer a generalisation in which dialogue can be treated as a special case of rational behaviour. The primary elements are accounts of planning and plan recognition, employing inference rules, action definitions, models of the mental states of the participants, and expectations of likely goals and actions in the context. The set of actions may include dialogue acts, whose execution affects the beliefs, goals, commitments, and intentions of the conversational partners.

2.5.6 Information State Update paradigm

The Information State Update (ISU) paradigm of dialogue modelling has emerged as a general framework for modelling flexible dialogue, see Poesio and Traum (1998); Traum et al. (1999); Bunt (1989; 2000); Larsson and Traum (2000). An 'information state' (also called 'context') is the totality of a dialogue participant's beliefs, assumptions, expectations, goals, preferences and other attitudes that may influence the participant's interpretation and generation of communicative behaviour (Bunt et al., 2010). The information state contains the information that a dialogue participant has at a given point during the dialogue, and every utterance in the dialogue leads to one or more information state updates. In other words, the effects of communicative acts (dialogue acts) are viewed as corresponding to update operations on the information states of understanding participants in the dialogue. The ISU-based approach allows dialogue modelling for various degrees of task complexity, and allows the dialogue system designers to decide about the level of dialogue flexibility to be tolerated: the differences are in the structure and contents of the information state, in the decision processes that use the information state as input, and in the update rules that manipulate it.

While an FSA defines all possible states of a dialogue in a finite set explicitly, the information state defines how the state of a dialogue may change, which, in general, makes it impossible to derive all reachable states since the number of states can be infinite. From the plan-based approach, the information state approach borrows the concepts of preconditions and effects of actions that change the state.

An assumption that is shared between all proposals for information states (e.g. Poesio and Traum, 1998; Bunt, 2000; Ahn, 2001; Cooper, 2004) is that an information state is structured into a number of distinct components. The information is, for example, divided into a 'private' part which contains *beliefs* which the participant assumes to be true; an *agenda* which contains short term goals or obligations of the agent; and a *plan* which contains actions or dialogue acts that the agent intends to carry out. A private part may also include 'temporary' shared information that has not yet been grounded, for instance including set of propositions that the participant *believes* to be true, a stack of *questions under discussion* (QUD), questions that have not been answered yet (see Ginzburg, 1998), and *latest utterance*, containing information about the latest utterance. The 'shared' part contains the same components as a 'temporary' shared one with the difference that this information has been grounded in dialogue, i.e. acknowledged by other participants. Figure 2.8 represents the information state of a dialogue participant as defined in (Traum et al., 1999).

Thus, the information state update approach consists of five key components: informational components, formal representations, dialogue acts (or moves), update rules and an update strategy.

1. The informational components specify what kinds of concepts are used to model the dialogue. A typical example is the internal state of a participant in a dialogue: for instance if a user drives a car and keeps closing his eyes, the system might realise that he is tired, which is then part of the internal state of the user. It is also typical to take external or environmental aspects into account, for example that the system will provide the tired user of the previous example with nearest hotels via GPS data. Moreover, it is useful to differentiate between static and dynamic contexts. Static context is information that remains fixed during the dialogue, for example that interaction is in English between two participants via a car control interface. Whereas dynamic context may change during the dialogue and from dialogue to dialogue, e.g. calendar entries in a smartphone that differ from user to user and constantly change. In addition, there is a wide range of further concepts that can be relevant, for example combinations of internal and external factors influencing the interaction.

2. The formal representation component deals with the question of how the specified concepts can be represented and used. The choices here range from simple data types, in particular numbers, strings, sets, stacks, lists, queues, but also more complex ones like records or typed feature structures, or complex information systems like higher-order logical propositions.

3. The third component is concerned with dialogue acts, see Section 2.1. They are interpretations of participants' dialogue intentions serving as update operators on the participants' information state. The DIT^{++} dialogue act annotation scheme (Bunt et al., 2010) and its subset - the ISO 24617-2 tagset provide a hierarchical multidimensional dialogue act taxonomy that has been used successfully for modelling dialogues of various complexities, genres and domains, see Section 2.6.

(1) U-Rule: **SelectAskCity**
PRE: fst (private.Agenda, raise(destCity(X))
EFF: set (NEXT-MOVE, ask(destCity(X))

(2) U-Rule: **IntegrateSysAskCity**
PRE: { val (shared.LM, ask(destCity(X))),
fst (private.Agenda, raise(destCity(X)))
EFF: { pop (private.Agenda),
push (shared.QUD, destCity(X))

(3) U-Rule: **IntegrateUsrAnswerCity**
PRE: { val (shared.LM, answer(destCity(X))),
fst (shared.QUD, destCity(X))
EFF: add (shared.BEL, destCity(X))

(4) U-Rule: **DowndateQUDCity**
PRE: { fst (shared.QUD, destCity(X)),
in (shared.BEL, destCity(X))
EFF: pop (shared.QUD)

(a) The update rules.

IS: [private: [BEL: { } ; Agenda: <raise(destCity(X)), ... >] ; shared: [BEL: { } ; QUD: < > ; LM: . . .]]

U-Rule: **SelectAskCity**
EFF: set (NEXT-MOVE, ask(destCity(X))

System: Where do you want to go?

U-Rule: **IntegrateSysAskCity**
EFF: { pop (private.Agenda)
push (shared.QUD, destCity(X))

IS: [private: [BEL: { } ; Agenda: < ... >] ; shared: [BEL: { } ; QUD: < destCity(X) > ; LM: ask(destCity(X))]]

User: Berlin

U-Rule: **IntegrateUsrAnswerCity**
EFF: add (shared.BEL, destCity(Berlin))

U-Rule: **DowndateQUDCity**
EFF: pop (shared.QUD)

IS: [private: [BEL: { } ; Agenda: < ... >] ; shared: [BEL: {destCity(Berlin)} ; QUD: < > ; LM: answer(destCity(Berlin))]]

(b) The example dialogue.

Figure 2.9: The update rules (a) and an example dialogue (b) between a system and a user illustrate the changes to the information state, based on Traum and Larsson (2003).

4. Update rules specify when and how an information state is to be updated and modified with the information specified in dialogue acts. They have three components: a name as identifier, a list of preconditions that must hold before a rule may be applied and a list of effects that specifies the changes that affect the information state once the rule is applied. Figure 2.9b illustrates a question-answer example based on the update rules from Figure 2.9a and the information state from Figure 2.8 to demonstrate how the information state is updated in this sample dialogue. Firstly, the agenda 'raise(destCity(X))' (asking the user about his destination city) fulfills the preconditions of rule (1) that selects a next move, namely 'ask'. After posing the question, the last move is set to the respective 'ask' move and the rule (2) is applied. This rule will add the item 'destCity(X)' to the QUD stack and at the same time delete this item from the private agenda. After the user's answer, rule (3) verifies whether his utterance is a valid answer as a destination city and integrates his answer into the shared beliefs of the information state. In a last step, the QUD is popped since the system has dealt successfully with the top item.

Several dialogue systems have been developed using the ISU framework, such as GoDIS (Larsson et al., 2000), IBiS1 (Larsson, 2002) and DIPPER (Bos et al., 2003).

2.5.7 (Partially Observable) Markov Decision Processes

For some domains, in particular constrained information seeking dialogues, statistical methods such as the (Partially Observable) Markov Decision Processes ((PO)MDP) dialogue systems have been shown to outperform rule- or knowledge-based dialogue models, see e.g. Lemon et al., 2006 for obtained evaluation results. In this approach, dialogue systems are modeled in the MDP framework and their policies are learned through statistical methods such as Reinforcement Learning (RL) [Sutton and Barto, 1998], see also Section 3.3. This has become a preferable approach to hand-crafted policies, since it can be easier optimised.

A very simple way to model dialogue states and transitions between them is a Markov Chain (Figure 2.10a). A Markov chain is defined by a set of states S and a transition model T specifying the probability $P(s'/s)$ of moving to state s' after being in state s at the prior time step. The state consists of the user intention, the last user action and the interaction context. However, this model does not tell anything about what actions the system should perform.

If the system performs an action, for example provides the requested information, there are two important things that happen. Firstly, the dialogue state changes depending on the action. Secondly, a system action is evaluated as being good or bad. System actions are influenced by the next state of the dialogue as well as by rewards for those actions (see Figure 2.10b). The resulting model is called a Markov Decision Process (MDP). Formally, it consists of:

- a set of states S
- a transition model $T : (S,A) \rightarrow S$

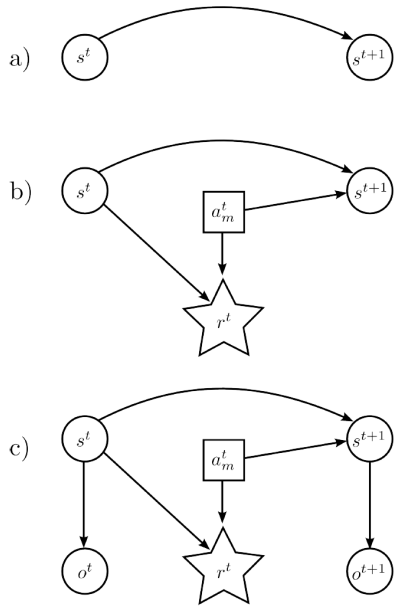

Figure 2.10: Diagram of state changes for different models. a) is a Markov chain, where s^t denotes the state at time t; b) is a MDP, where a_m^t is the system action and r^t the reward at time t; c) is a POMDP, where o^t denotes the observation at time t.

- a reward function $R : (S,A) \rightarrow \mathbb{R}$

The rewards in a MDP depend on the current system action and state. They can assume any real value. Generally, rewards are considered being negative for undesired system actions and positive for desired ones. Consequently, the agent's objective is to choose actions maximising his expected cumulative reward. Reinforcement learning is then applied to learn optimal dialogue policies. Its advantage is that the best sequence of actions does not need to be known in advance to be able to train the model. However, a pretty strong intuition is required to know which actions are good in which state.

Markov Decision Processes make a very unrealistic assumption. They require that the state of the dialogue can be directly observed by the agent. This is in conflict with many scenarios, in which (among others) the user's intention is part of a dialogue state. Often, an agent does not have direct access to the user's intentions. However, the agent might be able to infer the user's intention by exploiting the relation in which it is with directly observable variables. For example, the directly observable variable is the output of a speech recogniser processing the current user utterance. To integrate the concept of partial observability into the model, observations are made dependent on the state, which is hidden. This step results in a Partially Observable Markov Decision Process (POMDP). Formally, two things are added to the MDP model: (1) the set of observations O, and (2) the observation model Z which is defined as $P(o|s)$. As the states s^t are not directly observable in this model, the agent maintains a belief state, which is a distribution over the set of states S. The agent computes its belief state for the current time-step b' using its belief state b and action a_m at the previous time-step as well as its observation o at the current time step:

$$b'(s') = P(s'|b, a_m, o) = \alpha P(o/s') \sum_s P(s'|s, a_m) b(s)$$

The first equality can be justified by conditional independence and by replacing the states s^t in Figure 2.10c by belief states b^t. For example, we want to estimate the b^{t+1}. As the agent knows b^t and a_m^t, they are in the conditioning set. Rewards as well as previous belief states, actions and observations do not need to be considered because the flow of information to these is cut off due to s^t and a_m^t being forks on the corresponding paths. The second equality follows from factorising $P(s'|b, a_m, o)$. We observe, that the right hand side consists of an observation model $P(o|s')$ and a transition model $P(s'|s, a_m)$.

The (PO-)MDP approach has been successfully applied in ISU based systems, e.g. DIPPER-MDP and DIPPER-POMDP [Crook et al., 2010]. However, these approaches are used for rather simple closed domains, e.g. for a slot filling hotel booking task. To train such models large data amounts are typically required. It is harder to apply this approach at the level of more complex actions and plans, and for more complex interactive tasks such as, for example, natural argumentation dialogues, i.e. not slot filling tasks. Furthermore, to make Reinforcement Learning tractable, the state and action space must be carefully designed (Young et al., 2013;), which may restrict the expressive power and learnability of the model. Also, the reward functions needed to train such models are difficult to define and hard to validate in real time (Su et al., 2016).

2.5.8 End-to-end dialogue systems

Very recently, so-called end-to-end dialogue models have gained attention. Such systems are based on neural networks (Shang et al., 2015; Dodge et al., 2015; Serban et a., 2016; Wen et al., 2016). System components are directly trained on past dialogues without taking domain or dialogue structure into account. The approach escapes any semantic annotation and explicit modelling. Such systems typically do not include a dialogue management component. The approach is originally inspired by the sequence-to-sequence learning (Sutskever et al., 2014) to build end-to-end trainable, non-task-oriented conversational systems. Dialogue is treated as a source to target sequence transduction problem, applying an encoder network to encode a user query into a distributed vector representing its semantics, which then conditions a decoder network to generate each system response. The task is defined similar to a machine translation problem. Many types of neural networks have been applied to build end-to-end dialogue systems. A recent class of models are Memory Networks, see e.g. Sukhbaatar et al., 2015. The network is first writing and then iteratively reading form a memory component which can store entire past dialogues and short-term (local) context to compute the required response.

To achieve reasonable performance, training requires an enormous amount of data. Such systems mostly produce behaviour that is observed in data. They allow creating effective chatbot type systems but lack any capability for supporting domain specific tasks, for example, task-oriented information seeking dialogues, although efforts are made in this direction as well. For example, an approach proposed by Wen et al. (2015) defines each

system module as end-to-end trainable from data and a domain database operator. It also uses delexicalisation[4] and a weight tying strategy (Henderson et al., 2014c) to reduce the data required to train the model.

End-to-end systems are mostly evaluated using BLEU scores [Papineni et al., 2002]. It has been shown, however, that BLEU scores (or other word-overlap similarity metrics such as METEOR and ROUGE) do not correlate well with human judgments of dialogue quality, see Liu et al., 2016 and Georgilla et al., 2018.

2.6 Dynamic Interpretation Theory

Our approach is based on Dynamic Interpretation Theory (DIT) (Bunt, 2000). DIT has emerged from the study of spoken human-human information dialogues, with the aim of uncovering fundamental principles observed that can be utilised in the design of human-computer dialogue systems. In DIT, a dialogue is modelled as a sequence of utterances expressing sets of dialogue acts. Dialogue acts are semantic units, operating on the information states of the participants. A participant's information state in DIT is represented in a 'context model', which contains all information considered relevant for the interpretation and generation of dialogue acts. In updating the context model on the basis of dialogue acts, their preconditions and intended effects form the basis for changing the human participants' belief models. To compute the update semantics of various dialogue acts, an open multidimensional hierarchical dialogue act taxonomy was designed - DIT $^{++}$ Release 5[5]). The DIT $^{++}$ has a well-developed theoretical and empirical background, and served as the basis for designing the ISO 24617-2 dialogue act standard ([Bunt et al., 2010]; [Bunt et al., 2012a]; ISO, 2012).

The semantic DIT framework takes a multidimensional view on dialogue, in the sense that it views participation in a dialogue as being engaged in several activities simultaneously, such as trying to advance a task that motivates the dialogue, providing communicative feedback, taking turns, and so on. Communicative behaviour is interpreted in terms of bundles of update operations on participants' information states (or 'contexts'); such update operations consist of a semantic (referential, propositional, or action-related) content and a communicative function, which specifies what an addressee is supposed to do with the semantic content in order to update his information state [Bunt, 2007].

In DIT^{++}, the information which can be addressed is divided into: the domain or task (*Task*), feedback on the processing of previous utterances by the speaker (*Auto-feedback*) or by other interlocutors (*Allo-feedback*), managing difficulties in the speaker's utterance production (*Own-Communication Management*) or that of other interlocutors (*Partner Communication Management*), the speaker's need for time to continue the dialogue (*Time Management*), establishing and maintaining contact (*Contact Management*), the allocation of the next turn (*Turn Management*), the way the speaker is planning to structure the dialogue

[4]Delexicalistion replaces slots and values by generic tokens (e.g. keywords like Chinese or Indian are replaced by <v.food>) to allow weight sharing.

[5]https://dit.uvt.nl/

(*Dialogue Structuring*), and attention for social aspects of the interaction (*Social Obligations Management*).

It was observed in DIT^{++} that some utterances have communicative functions that can be applied to any kind of semantic content (*general-purpose (GP) functions*). In particular, they can be applied not only to content information concerning a certain task or domain, but also to information concerning the communication, e.g. an Inform like *'First of all we need to discuss the project finances'* is used to introduce a new topic into the discussion. *Dimension-specific* (DS) functions, by contrast, are applicable only to information concerned with a specific dimension of communication, e.g. using the utterance *'Let me see'* the speaker indicates that he needs some time to do something in preparation of continuing the dialogue (Stalling act). The phenomenon of general-purpose functions means that, when a stretch of communicative behaviour has a GP function, its full functional characterisation requires in addition also the specification of the dimension that is addressed, so we get characterisations like Auto-Feedback Check Question, i.e. $\langle autoFeedback; checkQuestion \rangle$, and Task Suggestion, i.e. $\langle Task; suggest \rangle$.

The ISO 24617-2 taxonomy is a subset of the DIT^{++} taxonomy. ISO 24617-2 has nine dimensions, Contact Management is considered as optional. Feedback functions corresponding to different processing levels are discarded and only positive, negative and elicitation functions are considered. Similarly, Discourse Structuring acts are reduced to two acts for Interaction management and Opening of a dialogue. Table 2.1 lists the 56 communicative functions defined in ISO 24617-2.

DIT^{++} and ISO 24617-2 can be extended with additional dimensions and communicative functions within each dimension, see Section 12 of the standard. For our application domains - debates and negotiations - we proposed extensions introduced in Chapter 4. We also completed the dialogue act update semantics with representations of domain-specific semantics for debates and negotiations, see Section 5.3.

Dialogue acts have a formally defined semantics. DIT is a context-change approach to dialogue acts and consider utterance meaning as defined in terms of how they affect the context. DIT models communicative agents as structures of goals, beliefs, preferences, expectations, and other types of information, plus memory and processing capabilities such as perception, reasoning, understanding, planning, etc. DIT uses the information state update machinery to tracking and understanding of the participants dialogue behaviour. For this purpose the DIT model provides a detailed specification of the creation, maintenance and use of shared beliefs. The model also provides procedures for incorporating beliefs and expectations shared between speaker and addressees in the tracking model. We will present details and examples in Section 5.6.

DIT^{++} and ISO 24617-2 are successfully applied to model dialogues of various complexities, genres and domains. Originally designed to model two-party human-computer dialogues where the system plays a role of an interactive cooperative assistant, e.g. to operate an electron microscope with the DenK system [Bunt et al., 1995], fax and copy machines using the DIAMOND system [Geertzen et al., 2004] and to interact with different graphical human-computer interfaces, e.g. the IDUSI system [Terken et al., 2006]. The DIT

General-Purpose	Dimension-Specific Communicative Functions	
Communicative Functions	Function	Dimension
Inform	AutoPositive	Auto-Feedback
Agreement	AutoNegative	
Disagreement	AlloPositive	Allo-Feedback
Correction	AlloNegative	
Answer	FeedbackElicitation	
Confirm	Staling	Time Management
Disconfirm	Pausing	
Question	Turn Take	Turn Management
Set-Question	Turn Grab	
Propositional Question	Turn Accept	
Choice-Question	Turn Keep	
Check-Question	Turn Give	
Offer	Turn Release	
Address Offer	Self-Correction	Own Communication Management
Accept Offer	Self-Error	
Decline Offer	Retraction	
Promise	Completion	Partner Communication Management
Request	Correct Misspeaking	
Address Request	Interaction Structuring	Discourse Structuring
Accept Request	Opening	
Decline Request	Init-Greeting	Social Obligations Management
Suggest	Return Greeting	
Address Suggest	Init-Self-Introduction	
Accept Suggest	Return Self-Introduction	
Decline Suggest	Apology	
Instruct	Accept Apology	
	Thanking	
	Accept Thanking	
	Init-Goodbye	
	Return Goodbye	

Table 2.1: ISO 24617-2 communicative functions.

semantic framework has been applied to model multimodal question answering interactions with medical information and healthcare services, e.g. the IMIX system (van den Bosch and Bouma, 2011) to design the PARADIME dialogue management system (Keizer et al., 2011). More recently, complex social interactions like multi-party meetings and games were modelled using the DIT^{++} and ISO 24617-2 dialogue act taxonomies (Petukhova, 2011, 2014 and Petukhova et al., 2015). For an overview of the use of the ISO 24617-2 standard see also (Bunt et al., 2017). The Cognitive Tutoring Systems to train metacognitive skills within political debate and negotiation contexts were designed applying the DIT theoretical framework and the DIT^{++} and ISO 24617-2 taxonomies to computationally model debate, negotiation and tutoring actions and strategies, and are discussed in this thesis.

2.7 Summary

In this chapter four main issues in dialogue analysis and dialogue modelling have been reviewed: aspects, qualities, facets of participating in dialogue, and aspects of involvement in dialogue.

Fundamental aspects of dialogue communication are concerned with the use of particular communicative acts in order to signal the speaker's state of beliefs, disbeliefs, and other attitudes, and general principles allowing the interpreter to reconstruct the relevant aspects of the speaker's cognitive state. These principles and their application in the interpretation and generation of specific kinds of communicative act form a basis for constructing and updating articulate dialogue models.

The meaning of dialogue units appropriate for computational modelling can be characterised in terms of communicative acts. A communicative act can be defined using three main concepts: intention (or purpose), effects and context. A communicative act has a purpose and has certain effects on the addressee. The interpretation of intention and effects is context-dependent. Adequate characterisation and formalisation of communicative act semantics in terms of intended context-changing effects on participants' information state is an important step forward in the analysis of dialogue phenomena, in the description of the interpretation of communicative behaviour of dialogue participants, and in the design of dialogue systems. Such a characterisation and formalisation is provided by the notion of a 'dialogue act' (Bunt, 1989) seen as an update operator on information states, and having two main components: communicative function and semantic content. Thus, describing communicative behaviour in terms of dialogue acts is a way of characterising the meaning of the dialogue behaviour, and the ultimate goal is to reconstruct the agent's intentions from the observation of his behaviour.

A phenomenon of fundamental importance is that dialogue contributions are often multifunctional. Dialogue participants have many (often parallel) tasks to perform during interaction. They not only need to obtain certain information, instruct another participant, negotiate an agreement, discuss results or plan future actions, etc. Among other things, dialogue participants have constantly to evaluate whether and how they can (and/or wish to) continue, perceive, understand and react to each other's intentions. They share information

about the processing of each other's messages, elicit feedback, monitor contact and attention, etc. Moreover, evidence suggests that understanding involves *parallel* processing and generation of multiple hypotheses. In human processing, all possible hypotheses are activated in parallel until it is possible to identify a single candidate or reduce their number. The hypotheses space when computing information from all processing modules in parallel is potentially very large. These observations have consequences for the overall dialogue system design since it may require additional strategies in the monitoring and use of resources (e.g. time and memory), and may impact processing and generation strategies. Multifunctionality of dialogue contributions needs to be first recognised as accurately as possible including functions that are not explicitly expressed in observed dialogue behaviour but are implied, entailed or used by default (Bunt 2007; Petukhova and Bunt 2010l and Petukhova 2011), and decisions need to be made with which hypotheses to update the system's information state, see Amanova et al., 2016 and Ebhotemhen et al. 2017 for meta-classification methods to manage the search and decision-taking process.

A full-blown dialogue model has to take the contribution in multiple modalities into account, as well as their integration (fusion). Many researchers increasingly propose that human cognition and human intelligent behaviour rely heavily on the modalities that constitute perception, action and mental states [Kiefer et al., 2008]. It is often relies on bodily states and physical actions [Niedenthal et al., 2005], is dependent on environment [Clark, 1998] and is situated in a social context [Barsalou et al., 2003, Tomasello, 2009]. In all situations, humans produce a continuous stream of perceptual experiences about respective situational content, along with corresponding conceptual interpretations of these situations (settings, involved agents and objects) which are based on previously experienced situations. We discussed the role of perceptual information related to motivation, affects, mental states, processing and generation of social signals for interactive learning situations modelled in this thesis.

To be involved in any interactive situation, dialogue participants need to generate useful (not necessary perfectly correct) inferences about setting, other agents and their actions, about objects, events, environment and possible actions. Therefore, participants' information states could be rather complex constructs accounting ideally for all relevant cognitive, affective and behavioural aspects. The better these aspects are modelled in a dialogue system the more adequate and rich behaviour it may generate. Dialogue participants do not only generate inferences based on their own 'private' beliefs and goals, but take their partners into account constructing 'shared' representations of the interactive situation. Both partners cooperate in the co-construction of the inferences, and they coordinate these actions. However, increasing complexity of the information states could affect system stability properties negatively due to the increase of the behavioural entropy, e.g. the entropy index has been shown to fluctuate with changes in the complexity of human information processing (see e.g. [Ramanand et al., 2003]) and decrease the reliability of information processing and learning, and task performance. Increasing complexity increases the number of variances in relationships between the states. When maximising the information diversity and richness, control and regulation should be also optimised. This can be achieved through adequate modelling and simulation. Well-defined grounding mechanisms help modelling expecta-

tions, projecting possible dialogue continuations and outcomes. They also enable adequate, flexible and yet computationally tractable information state update processes explaining how state change/transitions work. Additionally, incorporated metacognitive capabilities in the form of the system's proactive cognitive control is needed to anticipate the future demands of tasks, and improve performance on several elementary tasks and task switching. We discuss these regulation and decision making strategies in the next chapter.

Subsequently, we reviewed a number of existing approaches to computational dialogue modelling: Dialogue Grammars, Finite-State Automata, the frame-based approach, plan-based approaches, BDI-agent models, the Information-State Update framework, (PO)MDP and sequence-to-sequence models. All of them except for the last one make use of the notion of dialogue act and many model dialogue as a states transition task in order to achieve certain goals, accomplish a plan or obtain a reward.

Finally, we introduced the semantic framework of Dynamic Interpretation Theory used in this thesis as the basis for elaborate multidimensional computational dialogue context modeling and multi-agent dialogue management to deal with the complexities of human natural multimodal dialogue. We will show further that this framework, augmented with aspects of human cognition related to multitasking, prediction, learning and adaptation offers a plausible account of intelligent interactive human behaviour.

Chapter 3

Cognitive modelling of human dialogue behaviour

This chapter introduces the cognitive modelling task by discussing the important theoretical and empirical aspects related to the agency, parallel information processing and multitasking, human learning and human predictive and adaptive communicative behaviour. We present existing cognitive models: the cognitive task analysis and computer simulations, i.e. interactive cognitive agents. The widely used predictive CTA-based models, e.g. Hierarchical Task Analysis and GOMS, as well as ACT-R, a theory of human cognition which builds on the theory of rational analysis, are discussed.

Introduction

Cognitive models have been used for decades to explain and model human intelligent behaviour, and have been successful in capturing a wide variety of phenomena across multiple domains such as decision making (Marewski and Link, 2014), memory (Nijboer et al.,2016), problem solving (Lee et al., 2015), attention (Gunzelmann et al., 2009), perception (Salvucci, 2001), task switching (Altmann and Gray, 2008), user models in tutoring applications (Ritter et al., 2007), and neuroimaging data interpretation (Borst and Anderson, 2015).

Cognitive models are specifications of the mental representations, operations and problem solving strategies that occur during the execution of various tasks, e.g. computer based tasks. These models may be a rather complex and can take several forms: from relatively general descriptions of the steps required to complete a certain task to sophisticated computer simulations of users performing this task.

The task analysis method, in which a task is described in terms of a hierarchy of operations and plans, has been used successfully to simulate human decision-making processes. Two main paradigm have been developed: Behavioural Task Analysis (BTA) and Cognitive Task Analysis (CTA). BTA focuses on the behaviors people perform while doing their jobs. Typically, these behaviors are documented as discrete tasks or procedures which individuals

must accomplish to successfully perform a job (Jonassen et al., 1989). Focusing only on behaviour, however, is likely to produce information that is not very useful in understanding, aiding and training tasks [Means, 1993]. The focus has been put on the cognitive aspects of tasks that are not accessible to direct observations. It has been proven, that cognitive activities underlie even apparently simple tasks [Norman, 1993]. A technique which has been developed to help analyse the higher level cognitive functioning required in tackling complex tasks is Cognitive Task Analysis (CTA).

Available sophisticated cognitive models produce detailed simulations of human (multi-)task performance and can be used to implement cognitive agents to play a role in a multi-agent setting. It is of chief importance that artificial agents exhibit plausible human behaviour, notably a human-like way of learning and interacting. This means that such an agent makes decisions and takes actions that humans might also make and take, but also that the agent is influenced by its experiences and builds up his knowledge (mental representations) of the people it interacts with. Thus, the agent should be able to (1) process and perform several actions related to the underlying dialogue task and the roles it should play, e.g. as a partner or as a tutor; (2) learn by collecting a variety of experiences, through instruction and feedback, and through monitoring and reasoning about its own behaviour and that of partners; and (3) adapt its interactive behaviour to a human dialogue partner's knowledge, intentions, preferences and competences.

This chapter is concerned with theoretical and empirical aspects of cognitive modelling of human dialogue behaviour. Since the scientific focus of this thesis is to investigate integration of cognitive agent technology into dialogue management, we first provide the definition of agency and discuss its core properties (Section 3.1). In our analysis we model human dialogue behaviour as not only governed by cognitive processes of how people perceive, process, store, and apply information, but as a social interactive process. Therefore, the characterisation of agency is adopted from social cognitive theory and its properties are linked to fundamental principles of human social interaction such as cooperativity, rationality, sociality and ethics. Further, we introduce and discuss the fundamental concepts that play a key role in this study as related to multitasking (Section 3.2), learning (Section 3.3) and adaptivity (Section 3.4). Finally, we discuss the cognitive task analysis models, e.g. the most widely used GOMS model, and present the ACT-R (Anderson, 2007) platform that was used to build cognitive agents focusing on its declarative memory system and activation mechanisms (Section 3.5). Section 3.6 summarises the chapter by presenting our main observations and conclusions concerning the core properties of cognitive models related to metacognitive processes to be incorporated into the dialogue management of an interactive cognitive tutoring system. We present the 4C-ID based model which serves as the blueprint of the instructional design for our system. The design incorporates the cognitive task analysis model of in-domain and metacognitive skills training and the cognitive agent that simulates human decision taking behaviour and is able to perform multiple tasks in parallel, predict others' knowledge and intentions, learn and adapt. The design specifies supportive and procedural information concerning the training domain, interventions and feedback strategies in multimodal human-system dialogue based on sensory and multimodal input and output.

3.1 Agency and principles of human communicative behaviour

Following decades of research, a number of models of human dialogue behaviour have been designed attempting to identify core determinants and define fundamental principles/mechanisms of such behaviour. Modern dialogue theory assumes that dialogue participants act as motivated, cooperative, rational and social *agents* (Allwood, 2000; Allwood et al, 2000; Bunt 1994, etc). Dialogue participants act as senders and receivers, each with their own knowledge, beliefs, motivations, intentions and goals.

Cognitive science aims at deep understanding of human intelligent behaviour examining the nature, the tasks, and the functions of cognition. In its short history, cognitive theories have undergone several shifts. Initially, psychological input-output models were dominating assuming that human behaviour is shaped and controlled automatically and mechanically by environmental stimuli. Humans were not granted agentic capabilities. These models were later replaced by models of the mind as a 'digital computer' with the main assumption that information is fed through a central processor activating a solution according to predefined rules. Later, dynamically organised computational models were proposed performing multiple operations simultaneously and interactively to mimic better how the human brain works. These models include multilevel neural networks with intentional functions lodged in a subpersonal executive network operating without any consciousness via lower subsystems. Sensory organs deliver up information to a neural network acting as the mental machinery that does the constructing, planning, motivating and regulating non-consciously. Although more cognitive, these models still missed consciousness and agent capabilities.

Consciousness is the very substance of mental life. Consciousness plays a central role in the cognitive regulation of action and the flow of mental events (Carlson, 1997). People are *agents of experiences* rather that simply undergoers of experiences. Cognitive agents regulate their actions by cognitive downward causation as well as undergo upward activation by sensory stimulation (Sperry, 1993).

Modern cognitive theories merges two lines of research: microanalyses of the inner workings of the mind performing several cognitive tasks and macroanalysis of socially situated human development, adaptation and change. The core features of human agency are intentionality, forethought (or anticipation and planning), self-regulation by self-reactive influence and self-reflectiveness about one's capabilities.

To be an agent is to intentionally make things happen by one's action. Allwood et al (2000) characterise 'agenthood' by the following two principles:

- the intentionally controllable behavior of an agent is intentional and purposeful;
- the actions of an agent are not performed against his own will.

An intention is a representation of a future course of action to be performed (Bandura, 2001a). It is not a simple expectation or prediction of future actions but a proactive commitment to perform them. Outcomes are not the characteristics of agentive acts; they are the consequences of them. Actions intended to serve a certain purpose can cause quite different things to happen (Davidson, 1971).

Successful implementation of intentions requires certain self-regulatory aspects. People set goals anticipating the likely consequences of prospective actions, and select and create courses of action likely to produce desired outcomes and avoid undesired ones. Thus, behaviour is motivated and directed by projected goals and anticipated outcomes, i.e. 'anticipatory self-guidance'. In communicative situations, people communicate with the aim to achieve something (underlying task or activity) and they do this in a rational fashion (Bunt, 1994), organising the interaction so as to optimise the conditions for successful communication. The actions of a rational agent are performed only if the agent thinks it is possible to achieve their intended purpose (Allwood et al., 2000).

An agent is not only a planner and fore-thinker, but also a self-regulator. People constantly evaluate their behaviour and its outcomes. This includes self-monitoring, self-guidance and corrective self-reactions (Bandura, 1991). Thus, monitoring one's patterns of behaviour, the cognitive and the environmental conditions under which it occurs give rise to actions motivated primarily by personal goals and standards. Successful communication involves understanding perceptions, cognitive, meta-cognitive and emotional processes. The awareness and monitoring of one's own mental states and processes along with comprehension of other people's mental states is considered a factor of social adaptation both internal and external: high self- and other-monitors are more concerned that their interactions will go well, they are more likely to act to ensure this outcome, they also plan their actions more carefully, and they are able to flexibly modify their actions within the interaction in order to better adapt to the changing dynamics of the situation, typically by using other people's behaviour as a guide to their own (Ickes et al., 2006). People often take each others actions, motivations and other mental attitudes into consideration when acting, particularly for tasks such as dialogue.

Human functioning is rooted in a social system. 'Personal agency' operates within a broad network of socio-structural influences: rules, norms, social practices and sanction designed to represent human affairs (Bandura, 2001a). Communication is based not only on the human ability to perform motivated and rational action, but also cooperative and social ones. This also implies to take others into cognitive consideration, i.e. attempting to perceive and understand another person's actions; to have mutual awareness of shared purpose, agreement made about purposes and antagonism involved in the purpose; and to take each other into ethical consideration - make possible for the others to act freely, help others to pursue his/her motives, make it possible for others to exercise rationality successfully, see e.g. Allwood et al (2000). Communicative partners in dialogue are assumed to act according to the norms and conventions for pleasant and comfortable interaction (Bunt, 1996). Social cognitive theory also extends personal agency to collective agency which is not only a product of the shared intentions, knowledge and skills of its members but also of the interactive, coordinated and synergetic dynamics of their transactions. In this view, personal factors in the form of cognitive, affective and biological events, behavioural patters, and environmental events all operate as interacting determinants that influence each other in both directions, see Figure 3.1.

Thus, dialogue participants need to monitor all processes, to make use of resources and strategies available, to connect and organise different types of information, test and

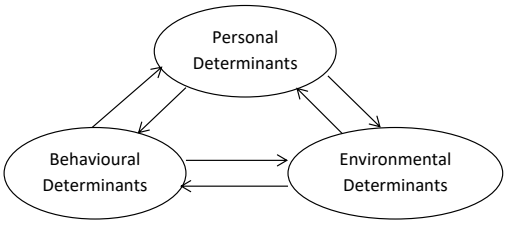

Figure 3.1: Triadic reciprocal causal model of social cognitive theory. From Bandura (2001).

modify, predict, and consequently plan and reason about future actions. This is possible if they act as motivated, rational, cooperative and social cognitive agents. Following these assumptions, we can find and explain regularities in dialogues by analysing (a) the relations between communicative devices in spoken language and nonverbal behavior, and the goals, purposes, and other circumstances on the part of a communicative agent that are expressed; (b) the system of communicative acts that can be realised with these communicative devices; (c) the internal structure of complexes of goals, beliefs, preferences, and other aspects of the communicative agent's state, which are revealed by its communicative acts.

3.2 Core tasks and roles of an agent

Every dialogue participant including artificial agents has three core tasks (at least): (1) to monitor partner dialogue behaviour; (2) to understand partner's dialogue contributions (i.e. intentions); and (3) to react adequately to partner's intentions by performing suitable actions to pursue a certain task or activity. Participants also need to share information about the processing of each other's messages, elicit feedback, manage the use of time, take turns, and monitor contact and attention. They often use linguistic and nonverbal elements in order to address several interactive and task-related aspects at the same time, as discussed in Section 2.2.

During interaction, an agent, including an artificial one, is mainly in the role of "speaker" (or "sender')' or in the role of "addressee" (also called "hearer" or "recipient"). It may also play the role of a side-participant who witnesses a dialogue without participating in it, see Clark (1996).

A dialogue system as an artificial agent has tasks dependent also on the application domain in relation to the role(-s) it plays, e.g. as a full-fledged interactive partner with equal responsibilities as a human one, as an assistant, adviser or mediator, as a passive observer, as a tutor or coach, and so on. For instance, for our applications of Virtual Debate and Negotiation Coaches we identified the following key roles:

- *Observer*: system observes dialogue sessions between two or more humans and keeps track of human-human dialogue without actively participating in it;

- *Mirror*: system re-plays the user's performance in a human-system dialogue in real time. The user observes his own performance and has the opportunity to terminate, re-enter and re-play the dialogue session from any point;
- *Experiencer*: system actively plays the role of one of the interaction participants, i.e. sender and addressee;
- *Tutor* or *Coach*: system provides feedback (corrective, verification, instructional, 'try again') from ongoing formative or summative assessment of user performance in one or more tutoring sessions [Mory, 2004].

The system may play multiple roles simultaneously and/or interchangeably.

In most existing approaches to dialogue management the Dialogue Manager (DM) is able to handle one particular dialogue task at a time. Most human activities however are essentially multitasking. For example, driving a car consists of two main processes: one that keeps the car in the middle of the driveway by looking at the road ahead of the car while operating the steering wheel and the gas and brake pedals, and a second process that monitors the traffic environment (e.g., is there a car behind you). Thus, human cognition can be conceptualised as a set of parallel cognitive modules (e.g. vision, declarative memory, working memory, procedural memory, manual control, vocal control, etc.). As long as multiple tasks do not need the same resources at the same time, these tasks can be carried out in parallel without interference. In the case of the driving example, if the driver is given an additional task, for example to operate a cell phone, he may abandon the monitoring task due to lack of resources.

Threaded cognition, as the theory of parallel execution of tasks, was proposed to explain human multi-tasking behaviour: why and when certain tasks may be performed together with ease, and which combinations pose a difficulty, what types of multitasking are disruptive, and when are they most disruptive. Threaded cognition models have been used in a wide spectrum of multi-tasking experiments, see Salvucci and Taatgen (2008), Salvucci and Taatgen (2010). This theory has been built on top of the ACT-R cognitive architecture. We designed a multi-threaded DM with integrated multi-tasking cognitive agent which, along with being an active dialogue participant with monitoring, understanding and reacting tasks, is capable of providing feedback on partner performance and which can reason about its own and partner's behaviour, and suggest alternative actions.

3.3 Human learning models for dialogue

Human learning involves acquiring new, or modifying and reinforcing existing knowledge, skills, values, and behaviours which may lead to a change in synthesising information, and the depth of knowledge, attitude or behavior relative to the type and range of experience (Gross, 2016). People learn from their own success and failures, from observing situations around them, and by imitating others' behavior. There are two widely used learning models: Reinforcement Learning (RL) and Instance-Based Learning (IBL).

Reinforcement learning (RL) is a formal model of action selection where the utility of different actions is learned by attending to the reward structure of the environment. Generally speaking, RL works in a trial-and-error fashion attempting various actions and recording the reward gained for those actions, see Sutton and Barto (1998). Reward can be associated with a specific strategy. One of the limitations of RL as a complete model of human decision-making becomes apparent in environments where goals change. This may happen due to changes in the environment or newly obtained knowledge of the environment. For example, you need to mail a letter, you looked for the closest post office in the neighborhood online, but on your way to it you see a street mailbox, so you drop the letter in there. Initial goal changes may occur due to the understanding and evaluation of partner behaviour. This often happens in negotiations where a negotiator may revise his initial offers and make concessions dependent on the interpretation of partner behaviour concerning these goals. RL models make decisions based solely on the learned state-action utilities. Rewards are set a priori, are fixed and never revisited. If the goal changes, the utilities representing the reward structure from the initial goal become irrelevant at best, and subversive at worst (Veksler et al., 2012).

Humans, by contrast, will employ their knowledge of the environment and about their interactive partners to make decisions for achieving new goals, for example, acting from experience or by association. People are adaptive, our memories are retrieved based on their recency and frequency of use (Anderson and Schooler, 1991) and strategies adapted with increasing task experience (Siegler and Stern, 1998).

Human learning often occurs as a result of experience. Decisions are made by finding a prior experience (an instance) that is similar to the current situation, see Logan (1988), Gonzalez and Lebiere (2005). Similarly, an agent can be trained by giving it a set of instances (learning-by-instruction), which it can refine and/or augment in actual interaction (learning-by-doing and learning-by-feedback). Decisions based on past experiences are stored in memory, and the most active is retrieved. Activation is based on history (e.g. frequency and recency) and on similarity (e.g. how similar the instance is, given the context).

RL is a useful paradigm where the possible strategies are relatively clear. If the underlying interaction structure is very flexible, unclear or absent (i.e. hard to derive on the basis of the system's behaviour), IBL has advantages. For instance, whenever a new goal is given, the IBL model will employ its stored knowledge (instances) to make informed goal-directed decisions. It does not need to learn the reward structure through trial-and-error; rather, the decision what action will be performed is based on the computed activation level, e.g. similarity between a past experience and the given current goal. Moreover, feedback can be used in IBL to create an instance that contains the correct solution, i.e. the model will add an instance of another strategy, whereas the RL model will punish the strategies that lead to a wrong solution. Strategy selection, which is implicit in RL, is explicit in the IBL model which makes it particularly suitable for tutoring applications. Exploiting advantages of the ACT-R architecture, namely the partial matching component in the ACT-R activation function, the IBL model is robust to missing or partial information, e.g. when the agent does not have full access to the same information as his partner or when the agent's knowledge is limited at the interaction beginning. We refer to Section 5.4 for an elaborate discussion

of the IBL-based model and its ACT-R implementation.

3.4 Adaptive dialogue modelling

Interactive systems and interfaces tailored towards specific users have been demonstrated to outperform traditional systems in usability. Nass et al. (2005) present an in-car user study with a "virtual passenger". Experimental results indicate that subjective and objective criteria, such as driving quality, improve when the system adapts its voice characteristics to the driver's emotion. Nass and Li (2000) confirm in a spoken dialogue study in a book shop that similarity attraction is important for personality expression: matching the users' degree of extroversion strongly influenced trust and attributed intelligence.

These observations have triggered the development of interactive systems that model and react to the users' traits and states in a timely fashion, for example by adapting the interaction based on language generation techniques (Mairesse and Walker, 2005). In Gnjatovic and Rösner (2008) a gaming interface is based on emotional states computed from the interaction history and actual user command. Nasoz and Lisetti (2007) describe a user modeling approach for an intelligent driving assistant that derives the best system action in terms of driving safety, given estimated driver states.

The above approaches adapt locally, i.e. the adaptation decision is made at turn level with very limited context and thus with no or very limited foresight. Reinforcement Learning (RL) has emerged as a promising approach for long-term considerations. While early studies (Walker et al., 1998; Singh et al., 2002) used RL to build strategies for simple systems; more complex paradigms are represented by statistical models, see Frampton and Lemon (2009). However, whenever users with different personalities in different states are systematically confronted with a learning system, most studies resort to user simulation: Janarthanam and Lemon (2009) simulate users of different levels of expertise, López-Cózar et al. (2009) simulate users with different levels of cooperativeness, and Georgila et al. (2010) simulate interactions of old and young users.

These studies demonstrate that the simulation of different user types is expected to lead to strategies which adapt to each user type. However, adaptivity has been not achieved at the level of dynamically changing goals within one dialogue. Rewards that are used in dialogue policy learning and optimisations are set a priori and are fixed. Human learning does not only involve strengthening of existing knowledge, compilation of new rules, collection of episodic experiences to improve future decisions, etc., but often requires more explicit reasoning, assessing why a particular solution worked or not, and manipulating the task representation accordingly - this is called 'metacognition'. In this study, metacognition plays two major roles: (1) it guides and regulates system task behaviour; and (2) it improves a participant's learning by triggering reasoning about one's own and partner behaviour.

Metacognitive skills can be trained by humans and learned by a system. When learning, humans also observe their partners' behaviour. In addition to using experiences to determine its own decisions, the agent can use them to interpret and reason about the behavior of others (i.e. humans). The ability to understand that other people have mental states, with

desires, beliefs and intentions, which can be different from one's own, is called Theory of Mind (ToM; Premack and Woodruff, 1978). Agents can learn about their partners' background knowledge, intentions and preferences, and respond to partner behaviour adequately and adapt their interactive behaviour to that of their partners. In our application, the ToM methodology has been used to infer, explain, predict and correct a partners' negotiation behaviour and negotiation strategies.

3.5 Computational cognitive models

3.5.1 Cognitive Task Analysis

Task analysis plays an important role in studying and modelling human intelligent behaviour and the processes behind it. It has a long history and has been successfully applied in designing a variety of applications in mission and organizational design, job design, quality improvement, training, error prediction, human-computer interaction and system design. Initially, task analysis was focusing on behavioural analysis (Behavioral Task Analysis, BTA) collecting, abstracting, organising, and reporting information about what people *do* when performing a task. BTA uses mainly three different methods, where, (1) competent individuals who have demonstrated expertise, called subject-matter experts (SMEs) are observed and steps documented; (2) experts are consulted in performing the required task, e.g. in structured and semi-structured interviews; and (3) the actual users are observed performing the task themselves and steps are documented. [Mager, 1997] described task analysis as *a collection of techniques used to help make the components of competent performance visible*. In more detail, performing behavioral task analysis means observing performance, describing it in words, unpacking the description hierarchically into sub-procedures, continuing the process until some assumed elemental level of description is reached, identifying conditional antecedents and measurable outcomes for each element, and finally consolidating commonalities across the hierarchy.

Hierarchical Task Analysis

Beginning in the 1970s, researchers expanded on the importance of cognition in task analysis [Annett et al., 1971]. Cognitive Task Analysis (CTA) attempts to describe or analyse the mental phenomena that engender specific behaviors. The focus is on the mental representations, underlying knowledge, thought processes and goal structures that underlie observable task performance, i.e. cognitive processes and activities necessary to make decisions and perform actions (Chipman et al., 2000). CTA is typically carried out when knowledge about how a task is performed is uncertain.

It has been noticed that focusing primarily on simple observable behaviors makes it impossible to capture the dynamic non-linear nature of many jobs. Hierarchical Task Analysis (HTA) has therefore been proposed and applied in human-machine systems [Annett, 2003]. Modern HTA considers not only what should happen, but also predicts what can actually

happen and what can go wrong. HTA assists in discovering the success and failure indicators of each of sub-goals. This makes this method particular useful for our application - interactive cognitive tutoring systems. HTA aims at greater understanding of cognitive tasks. At the core is goal-directed behaviour comprising a sub-goal hierarchy linked by plans. The three main principles governing the analysis have been formulated as follows (Annett et al., 1971):

1. At the highest level a task is considered as consisting of an operation and the operation is defined in terms of its goal. The goal implies the objective of the system in some real terms of production units, quality or other criteria.

2. The operation can be broken down into sub-operations each defined by a sub-goal again measured in real terms by its contribution to overall system output or goal, and therefore measurable in terms of performance standards and criteria.

3. The important relationship between operations and sub-operations is really one of inclusion; it is a hierarchical relationship. Although tasks are often proceduralised, that is the sub-goals have to be attained in a sequence, this is by no means always the case.

Thus, HTA is proposed to describe a system in terms of its goals, which are expressed in terms of some objective criteria. Subsequently, HTA breaks down sub-operations into a hierarchy. Sub-operations are defined as sub-goals, sub-goals are described again in terms of measurable performance criteria. Finally, there is a hierarchical relationship between goals and sub-goals and there are rules guiding the sequence that the sub-goals are attained, see [Piso, 1981, Hodgkinson and Crawshaw, 1985, Bruseberg and Shepherd, 2017].

The flexibility of the HTA method, enabling a semi-structured approach (Figure 3.2) can be used to describe many aspects of an application, e.g. training requirements, error prediction, performance assessment and system design. This makes it attractive for our application as well. More recent innovations are various templates, e.g. sub-goal (Figures 3.3) and plan templates (Figure 3.4) that help to formalise the HTA processes [Ormerod and Shepherd, 2003].

Training design Piso (1981)	Interface design Hodgkinson and Crawshaw (1985)	Job design Bruseberg and Shepherd (1997)
What is the goal of the task?	What are the sensory inputs?	How does information flow in the task?
What information is used for the decision to act?	How can the display of information be improved?	When must tasks be done?
When and under what conditions does the person (system) decide to take action?	What are the information processing demands?	What is the temporal relation of tasks?
What are the sequence of operations that are carried out?	What kind of responses are required?	What are the physical constraints on tasks?
What are the consequences of action and what feedback is provided?	How can the control inputs be improved?	Where can and cannot error and delay be tolerated?
How often are tasks carried out?	What kind of feedback is given?	Where is workload unacceptable?
Who carries the tasks out?	How can the feedback be improved?	Where is working knowledge common to more than one task element?
What kinds of problems can occur?	How can the environmental characteristics be improved?	Where do different tasks share the same or similar skills?

Figure 3.2: Examples of questions for sub-goals definition for three different application domains when performing Hierarchical Task Analysis.

SGTs	Task element	Context for assigning SGT and task element
Act: To operate as part of a procedure	A1: Activate	To make a subunit operational, e.g., to switch from an 'off' state to an 'on' state
	A2: Adjust	To regulate the rate of operation of a unit maintaining an 'on' state
	A3: Deactivate	To make a subunit non-operations, e.g., to switch from an 'on' state to an 'off' state
Exchange: To exchange information	E1: Enter	To record a value in a specified location
	E2: Extract	To obtain a value of a specified parameter
Navigate: To search for information	N1: Locate	To find the location of a target value or control
	N2: Move	To go to a given location and search it
	N3: Explore	To browse through a set of locations and values
Monitor: To monitor system state and look for change	M1: Detect	To routinely compare the system state against the target state in order to determine the need for action
	M2: Anticipate	To compare the system state against the target state in order to determine readiness for a known action
	M3: Transition	To routinely compare the rate of change during a system state transition

Figure 3.3: Sub-goal templates (SGT) for Hierarchical Task Analysis. Adopted from Ormerod and Shepherd (2004).

Code	Plan Type	Syntax
S1	Fixed sequence	Do X, Y, Z
S2	Contingent sequence	If (c) then do X
		If not (c) then do Y
S3	Parallel sequence	Do together X, Y, Z
S4	Free sequence	In any order do X, Y, Z

Figure 3.4: Plan templates for Hierarchical Task Analysis. Adopted from Ormerod and Shepherd (2004).

Sensemaking models

As stated above, contemporary task analyses involve both behavioral and cognitive techniques. Behavioral task analysis focuses on the identifiable behavioral activity that a user needs to perform. Most practitioners recognise that monitoring, detecting, recognising, and deciding are essential components of any task analysis. Therefore, all successful task performance involves at least some cognitive components related to perception, decision making, knowledge, and judgment (Welford, 1968).

Task analysis may be seen as a form of *sensemaking* - the process of searching for a representation and encoding data in that representation to answer task-specific questions. Different operations during sensemaking require different cognitive and external resources. It may be assumed that experts will have built up from extensive experience a set of patterns around the important elements of their tasks, which we here call *expertise schemas*. Thus, the key to expert performance is to design domain-specific schemas (Ericsson and Lehmann, 1993). Experts do not just automatically extract patterns and retrieve their response directly from memory. Instead, they select the relevant information and encode it in special representations, i.e. schemas, that allow planning, evaluation and reasoning about alternative courses of action.

Tasks would consist of information gathering, organising this information in a schema that aids analysis, the development of insight through the manipulation of this schema, and the creation of some knowledge product or direct action based on the insight: Information $->$ Schema $->$ Insight $->$ Product.

The sensemaking process starts with search for data, typically large amounts. The raw data undergoes transformation as it is turned into information, e.g. to reportable results. External data sources are the raw evidence. A much smaller subset of that external data is relevant for further processing. Typically, evidence is selected and extracted. Subsequently constructed schemas are the re-representations of the information so that it can be used more easily to draw conclusions. Schemas are used to generate hypotheses, i.e. make predictions or gain insights, are the tentative representation of those conclusions with supporting arguments. Ultimately there is a product. Basically the data flow represents the transduction of information from its raw state into a form where expertise can apply and then out to another form suited for the task performance.

As result of the sensemaking process, a task model incorporates data from the research literature, expert knowledge and empirical evidences obtained from the interactions and

experiments with real users.

Goals, Operators, Methods, and Selection rules (GOMS)

The oldest and still most widely used approach to modelling human-computer interaction is based on a model of human information processing and a task analysis method proposed by Card et al. (1983). At the core of this approach is the GOMS task analysis method:

- Goals: specifications of user's goals that he aims to achieve in an interaction;
- Operators: the possible actions that an interface allows to be performed by a user, e.g. clicking, dragging with the mouse cursor, using speech, typing, selections on touch screen, etc.
- Methods: sequences of sub-goals and operators that can be used to achieve a certain goal. There is often more than one method to accomplish a goal;
- Selection rules: rules by which a user chooses a particular method from a number of alternatives for a achieving a goal.

The GOMS analysis is a predictive model used to predict several aspects of human performance when he interacts with an interface. Goals are broken down into sub-goals. All sub-goals must be accomplished in order to achieve the overall goal. The primary aim of the method is to specify goals, unit tasks and sub-goals. Goals and sub-goals are often arranged hierarchically, but a strict hierarchical goal structure is not required. For instance, some behaviour goals can be modeled as 'flattened' structures. In some cases several goals need to be active at once.

There are four basic GOMS modelling techniques: the Keystroke-Level Model (KLM, Card et al., 1980b); the Card, Moran, and Newell GOMS (CMN-GOMS, Card et al., 1980a); the Natural GOMS Language (NGOMSL, Kieras, 1988) model; and the Cognitive-Perceptual-Motor GOMS (CPM-GOMS, Card et al., 1983) model.

The Keystroke-Level Model is the simplest GOMS technique that does not include goals or selection rules but simply specifies the sequence of operators and methods required to perform a task. The main method goal is to predict the execution time of interactive tasks. Original heuristic rules were created primarily for command-based interfaces and updated for direct-manipulation interfaces.

The CMN-GOMS method presuppose a strict goal hierarchy. Methods are represented in an informal program form that can include sub-methods and conditionals. A CMN-GOMS model, given a particular task situation, can thus predict both operator sequence and execution time.

The NGOMSL model is in program form and provides predictions of operator sequence, execution time, and time to learn the methods. An NGOMSL model can be constructed by performing a top-down, breadth-first expansion of the user's top-level goals into methods, until the methods contain only primitive operators, e.g. keystroke-level operators. Like

CMN-GOMS, NGOMSL models explicitly represent the goal structure, and so they can represent high-level goals. The NGOMSL technique refines the basic GOMS concept by representing methods in terms of a cognitive architecture called *cognitive complexity theory* (CCT, Bovair et al. 1990). CCT assumes a simple serial-stage architecture in which working memory triggers production rules that apply at a fixed rate. These rules alter the contents of working memory or execute primitive external operators, e.g. making a keystroke. GOMS methods are represented by sets of production rules. Learning procedural knowledge consists of learning the individual production rules. Learning transfers from a different task if the rules had already been learned (see also Anderson, 1993).

The CPM-GOMS models predict execution time based on an analysis of component activities. In CPM-GOMS the primitive operators are simple perceptual, cognitive, and motor acts. Unlike the other extant GOMS techniques, CPM-GOMS does not presuppose that operators are performed serially; rather, perceptual, cognitive, and motor operators can be performed in parallel as the task demands. CPM-GOMS is based directly on the Model Human Processor (MHP) (see Card et al., 1983), which is a basic human information-processing architecture. The human is modeled by a set of processors and storage systems in which sensory information is first acquired, recognised, and deposited in working memory by perceptual processors, and then a cognitive processor acts upon the information and commands motor processors to make physical actions. Each processor operates serially internally, with a characteristic cycle time, but processors run in parallel with each other. There are templates provided that are assemblies of cognitive, perceptual, and motor operators and their dependencies defined for different activities under different task conditions.

Although GOMS approaches have been useful in providing a range of techniques for analysing interactive behaviour and predicting execution times for a variety of tasks, it has a number of constraints. First of all, GOMS based models can model only expert performance. The assumption is rather strong that users are well practiced, make no errors during the task or perform worse over repeated performance of a task due to fatigue. Secondly, it is difficult (and rather expensive) to model novel interaction methods, e.g. spoken human-computer interactions or multimodal ones. It is necessary to first conduct some small experiments to measure the average length of time it took to perform certain actions using the different interaction methods.

3.5.2 ACT-R cognitive architecture

Computational cognitive models are playing an increasingly important part in HCI research. Unlike the GOMS family of models which can model only expert performance, computational cognitive models allow researchers to show how users learn to use systems and also how users may perform in certain circumstances. Unlike artificial intelligence programmers who aim to build programs that complete tasks faster and with fewer errors than humans, computational cognitive modellers try to write computer programs that complete tasks in exactly the same way as humans i.e. that make all the mistakes that humans make, and take as much time as humans do to perform tasks.

ACT-R, Anderson, 2007, is the theory and platform for building models of human cog-

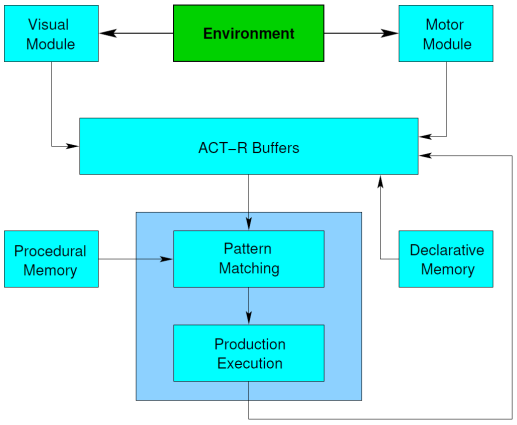

Figure 3.5: The modular structure of ACT-R 6.0. Adopted from Bothell, 2004

nition. ACT-R offers a platform for modelling human behaviour and has been applied successfully in the past for many tasks, in particular to model human decision making that is learnable, adaptive and evolving over time (Gonzalez et al., 2003). ACT-R accounts for hundreds of empirical results obtained in the field of experimental psychology and proposes a hybrid architecture that combines a production system to capture the sequential, symbolic structure of cognition, with a sub-symbolic, statistical layer to capture the adaptive nature of cognition. Computational cognitive models are models which help designers to understand how the human mind works and how users learn and interpret information and how they interact with computers.

The ACT-R architecture assumes a mixture of parallel and serial processes. The core ACT-R components are: the perceptual-motor system, the goal system, the declarative memory and the procedural system, see Figure 3.5.

The main model functionality relevant for our application comes from activation of declarative knowledge within ACT-R's declarative memory system. ACT-R's declarative memory for facts is one of the central notions and consists of a network of schematic units known as *chunks*. Each chunk has one or more slots that contain values or links to other chunks in declarative memory. Each chunk also possesses an activation value reflecting their use: chunks that were used recently and chunks that are used frequently get a high activation. More active chunks are more likely to be retrieved in a search of declarative memory. The activation level of a chunk (i) is determined by using an architectural mechanism incorporating the past history of a chunk i use and derived from the following equation, see [Bothell, 2004]:

$$A_i = ln(\sum_{j=1}^{n_i} t_{ij}^{-d}) + Logistic(0,s)$$

where n is the number of times an instance i has been retrieved in the past; t represents the amount of time that has passed since the j_{th} presentation or creation of the instance, and d is the rate of activation decay.[1] The rightmost term of the equation represents noise added

[1] In the ACT-R community, 0.5 has emerged as the default value for the parameter d over a large range of

to the activation level, where s controls the noise in the activation levels and is typically set at about 0.25, consistent with the value used in Lebiere et al. (2000). Thus, the equation effectively describes both the effects of recency - more recent memory traces are more likely to be retrieved, and frequency - if a memory trace has been created or retrieved more often in the past it has a higher likelihood of being retrieved.

An instance does not have to be a perfect match to a retrieval request to be activated. ACT-R can reduce its activation according to the following formula used to compute partial matching P_i, see [Bothell, 2004]:

$$P_i = \sum_l PM_{li}$$

where M_{li} indicates the similarity value between the relevant slot value in the retrieval request (l) and the corresponding slot instance i summed over all slot values in the retrieval request. P denotes the mismatch penalty and reflects the amount of weighting given to the matching, i.e. when P is higher, activation is more strongly affected by similarity. We set the P constant at 5, consistent with the value used in Lebiere et al. (2000).[2] Thus, the model can retrieve chunks that are not exact matches, but the mismatch penalty makes this less likely. This has two important implications for our model. First, it means that the model will be able to retrieve past instances for reasoning even when the model has not encountered a particular situation before. Partial matching, combined with activation noise, also allows for flexibility in the model's behaviour. It will not rigidly make the exact same moves every time.

3.6 Discussion and conclusions

In this chapter we discussed the main relevant aspects of the cognitive modelling of the human social interactive dialogue: characterisation of agency linked to fundamental principles of human communicative behaviour; key cognitive capabilities related to multi-tasking, abilities to learn, anticipate and exhibit adaptive interactive behaviour. We also introduced the Cognitive Task Analysis methodology and the ACT-R architecture, in particular its declarative memory system and activation mechanisms.

It has been observed by many researchers that human intelligent behaviour exhibits certain patterns and regularities, and is performed in accordance with norms and conventions applicable in a particular social interactive situation. The assumption that dialogue participants act as motivated, cooperative, rational and social agents allows to explain such regularities. This assumption is extremely useful for modelling the fundamental aspects of dialogue communication. Human agency involves intentionality, anticipation, self-regulation and self-reflection. Along with the agent's knowledge and understanding ability, important actions to achieve effective results are to control and manipulate one's cognitive processes: monitor the degree to which it understands the user's behaviour, obtain and apply new information, recognise failures, employ effective repair strategies, anticipate outcomes of its

applications, [Anderson et al., 2004].

[2] To disable partial matching P can be set at 0.

own and partner actions, adapt its behaviour (reactive and pro-active) to partner performance and needs, and learn effectively from these experiences.

Human dialogue behaviour is multitasking. In many interactive situations, people typically share and vary responsibilities in observing, monitoring, experiencing and executing different tasks, and become aware of different strategies and how they work.

Many learning systems require long training and large sets of examples. The Instanced-Based Learning (Lebiere et al., 1998), which is similar to human learning, can be successful with small numbers of training examples. Instance-based learning has been applied to provide explanations for skill acquisition (Lebiere and Anderson, 1998), categorisation (Anderson and Betz, 2001), human decision making (Gonzalez et al., 2003), language acquisition (Taatgen and Anderson, 2002), development of theory of mind (Arslan et al. 2017) and game playing (Meijering et al., 2014). The core idea of Instance-Based Learning is that examples of task performance are stored in a memory model, from which relevant examples or blendings of examples can be retrieved for future use. Instance-based learning has been implemented in the ACT-R (Anderson, 2007) cognitive architecture and simulates successfully the human decision making processes and other aspects related to human interactive performance.

Humans may dynamically change their initial goals; they often revise their previous decisions and adapt their task-related and communicative strategies. An agent that is flexible in anticipating, taking and revising decisions, will show adaptive behaviour enabling the dialogue system to communicate with users more naturally and efficiently. The agent will be adaptive in setting goals, proactive in choosing appropriate strategies and persistent in monitoring progress.

The assumptions, observations and specifications summarized here will enable building cognitive task models of the in-domain and metacognitive skills training and cognitive agents that are plausible conversational partners performing actions and making decisions similar to those that humans will do.

For our use case (see Section 4.3 for more detail) - interactive cognitive tutoring - we need an educationally sound model of (meta)cognitive processes behind task-related, debate and/or negotiation, skills development. Modern research takes the view that overt observable behaviour and covert cognitive functions behind it form an integrated whole. In educational design practice, there is a growing interest in using whole-task models. *Whole-task models* aim to assist learners in integrating knowledge, skills and attitudes into coherent wholes, to facilitate transfer of learning. Tutors, including artificial ones, are then informed about how to balance the load of the learner, make the tasks sufficiently challenging, and how to provide feedback. In particular, we refer to 4C-ID instructional design model (see e.g. Van Merriënboer and Kirschner (2013), van Rosmalen et al., (2015) and Section 4.3 of this thesis) that prescribes how to train complex cognitive skills. The key elements of this model are:

1. Authentic, whole tasks preferably based on real-life tasks and organised in task classes with variation and increasing complexity.
2. Supportive information to the non-recurrent aspects of the tasks and explanation how

a domain is organised. This information is always available.

3. Procedural information concerning recurrent aspects of tasks and instructions how to perform the routine aspects of a task. This information is available just-in-time and typically will fade out stepwise when exercising with new tasks.
4. Part-task practice: additional practice for routine aspects of learning tasks that require a high level of automation.

The model fits well the instructional design for our domains where the users have to stepwise understand and learn how to present and argue defending their positions. They work with realistic, engaging tasks adjusted to their personal needs in terms of complexity levels, and if necessary, have the option to practice selected types of sub-tasks, e.g. presentation skills in delivering convincing performance, core and advanced argumentation and rebuttal skills, argument structuring in such a way that presented evidence is acceptable and relevant, and sufficient to draw valid conclusions. To design tasks hierarchies, we used the core CTA techniques and requirements analysis methods, such as available multimodal training material, (semi)structured interviews, think aloud experiments, training guidelines and evaluation performance criteria, teaching methods and available statistics or best practices, etc. All these grounded in real-life examples. Expert and novice interactions performing the same or comparable tasks were video recorded and analysed, see e.g. [Petukhova et al., 2015c, Petukhova et al., 2017b]. The CTA analysis, more specifically the semi-structured HTA carried out, enabled us to identify the expectations, differences and overlaps in task performance and training methods. The analysis enabled rather detailed predictions of the accuracy and efforts required to execute practiced actions by learners. The model is used not only to detect errors but also to explain why learners' actions were incorrect. The CTA methods supported our design of the instructional models of metacognitive skills development for an intelligent tutor. The adoption of the 4C-ID model did help us to prepare the global design of the interactive cognitive tutoring application in an educationally sound way, see also [Van Rosmalen et al., 2015].

The designed cognitive agents build representations of the people they interact with, and modify their own behavior accordingly. The specified observations call in the first place for an articulate multi-dimensional dialogue context model that enables multiple simultaneous and independent updates, application of various update mechanisms, and monitoring and control processes. Agents operate on the basis of such a model situating them in concrete real interactive scenarios. We will show in the next three chapters that the application of cognitive models to build cognitive task agents has a potential to advance the design of dialogue system that show flexible adaptive but also robust dialogue behaviour with limited data resources.

The design of cognitive, instructional and dialogue models is informed by numerous data collection, analyses and evaluation experiments involving real human users presented in the next chapter.

Chapter 4

Data-driven dialogue system design

This chapter addresses the data-driven dialogue system design. The corpus development is performed within the ISO linguistic annotation framework and primary data encoding initiatives. The Continuous Dialogue Corpus Creation (D3C) methodology is proposed, where a corpus is used as a shared repository for analysis and modelling of interactive dialogue behaviour, and for implementation, integration and evaluation of dialogue system components. All these activities are supported by the use of the ISO standard data models including annotation schemes, encoding formats, tools, and architectures. Standards also facilitate practical work in dialogue system implementation, deployment, evaluation and re-training, and enabling automatic generation of adequate system behaviour from the data. The proposed methodology is applied to the data-driven design of two multimodal interactive applications - the Virtual Negotiation Coach, used for the training of metacognitive skills in a multi-issue bargaining setting, and the Virtual Debate Coach, used for the training of debate skills in political contexts. The chapter also presents an approach to achieve interoperability of dialogue act annotations through developing a query format for accessing existing annotated corpora. The interpretation of expressions in the query format implements a mapping from ISO 24617-2 concepts to those of several existing annotation schemes.

Introduction

A steadily growing interest can be observed in data-driven modelling of phenomena related to natural language, vision, behavioural and organisational processes. Data have become essential to advance the state of the art in many areas including the development of spoken (multimodal) dialogue systems. Many commercial conversational applications, e.g. Apple's Siri, Microsoft's Cortana and Google Now, became successful and robust partly due to the

Chapter is largely based on Malchanau et al.(2018b), for which I performed the research in close collaboration with my co-authors. Section 4.6.3 reports is based on research reported in [Petukhova et al., 2014b]. My specific contribution was the design of the ISO-based querying format and the corpora querying tool implementation.

amount of real user data available to their developers. The most recent trend in dialogue system design involves end-to-end dialogue systems that use neural network models trained on a large amount of dialogue data, without any detailed specification of dialogue states [Wen et al., 2017, Bayer et al., 2017].

Dialogue data have often been collected in Wizard-of-Oz experiments (Dahlbäck et al., 1993), where the dialogue system is replaced by a human Wizard who simulates the system's behaviour according to a pre-defined script.

An alternative is to use simulated users. With good user modelling, a dialogue system could be rapidly prototyped and evaluated. Simulated data sets are, however, rather scarce [Schatzmann et al., 2006].

Resources for data-driven learning of task-oriented systems are also collected with existing systems [Bennett and Rudnicky, 2002, Henderson et al., 2014a]. For example, the DialPort project addresses the need for dialogue resources by offering a portal connected to different existing dialogue systems [Lee et al., 2017].

Learning algorithms have also been proposed to train a dialogue system online. System behaviour is initially learned from a minimal number of dialogues and is then optimised as more data arrives [Daubigney et al., 2012]. As a data collection strategy this approach may not be really successful, since the initial system performance can be rather poor.

Building an annotated dialogue corpus is an expensive activity, especially when it requires manual annotations. Over the years, many annotated dialogue corpora have been created, however annotations and their formats differ from resource to resource. The community has recognised this problem by addressing the interoperability of dialogue resources. ISO 24617-2 "Semantic annotation framework, Part 2: Dialogue acts" [ISO, 2012], in particular aims to contribute to the interoperability of annotated dialogue corpora. New corpora have been created [Petukhova et al., 2014a], existing corpora re-annotated [Bunt et al., 2013] using the standard annotation scheme, and existing annotations mapped to ISO 24617-2 [Petukhova et al., 2014b]. The DialogBank is a new language resource that contains dialogues of various kind with gold standard annotations according to the ISO 24617-2 standard [Bunt et al., 2016].

We propose the *continuous dialogue corpus creation* (D3C) approach where the corpus is exploited as a shared repository for analysis and modelling of interactive dialogue behaviour, and for implementation, integration and evaluation of the dialogue system. The method situates the corpus development in the framework of ISO linguistic (i.e. semantic) annotation standards[1]. Standard data models (i.e. annotation schemes, encoding formats, tools, architectures) support the corpus development facilitating the creation of semantically rich and interoperable dialogue data for multiple domains, contributing to cost reduction in corpus creation and dialogue system design and re-training, and enabling automatic generation of adequate system behaviour from the data. The developed corpus is enriched with additional interoperable semantically annotated dialogue resources through *querying* various existing annotated corpora applying format whose expressions make use of the annotation language defined by the standard. The interpretation of the expressions in the query

[1] We refer to the work of Ide and Pustejovsky, 2017 for an overview of existing standards.

implements a mapping from ISO 24617-2 concepts to those of the annotation scheme used in the corpus.

The chapter is structured as follows. Section 4.1 presents the overall methodology discussing the main principles and key processes related to corpus development. Section 4.2 presents the ISO 26417-2 data model introducing the basic concepts and the Dialogue Act Markup Language (DiAML) as the main corpus annotation and exchange format between system components. Further, we illustrate the proposed D3C approach applying it to recently performed corpus activities when designing two different applications - Virtual Debate and Negotiation Coaches, Chapter 6. We discuss the intelligent tutoring use case and two scenarios - debate and negotiation settings for training metacognitive skills. Subsequently, we present data collection and processing. The in-domain data collection is augmented with out-of-domain dialogue data aggregation. The method of querying existing dialogue resources is presented. We define the DiAML-based query format used to query not-ISO 24617-2 annotated corpora for which the mapping between dialogue act concepts defined in the ISO standard and those of other dialogue act annotation schemes exist. The method is illustrated by querying two widely used dialogue resources; the AMI and MapTask corpora (4.6.3), and is evaluated with respect to accuracy and completeness through statistical comparisons between retrieved and manually annotated corpus data. We show how standards and interoperable semantic resources facilitate practical work in dialogue system implementation, deployment, evaluation and re-training, and enables automatic generation of adequate system behaviour from the data.

4.1 Continuous corpus creation methodology

An important step in designing any multimodal dialogue system is to model natural human dialogue behaviour, as a basis for developing dialogue system components. Each module in a dialogue system performs a task such as dialogue act classification, event identification, co-reference resolution, or semantic role labelling, and is integrated according to the adopted architectural approach (e.g. pipeline, multi-agent or multi-threaded), which determines how the modules communicate and exchange their processing results. Success in such a development will heavily depend on the *quality*, *costs* and *application range* of the underlying corpus data. These three aspects are influenced by multiple variables such as number, tasks and roles of dialogue participants involved in an interactive situation (real vs simulated humans vs artificial agents); dialogue setting, modalities and media available; granularity and nature of annotations and analysis (manual vs automatic vs no annotations); types of infrastructures, tools and data formats. All these variables impact the corpus creation design, the complexity of the set-up, and the processing steps.

We propose a continuous dialogue corpus creation (D3C) methodology consisting of the following steps (Fig. 4.1):

1. **Set-up**: structuring from the ISO 24617-2 metamodel, define an interaction scenario and specify data collection requirements; provide details for participants roles and

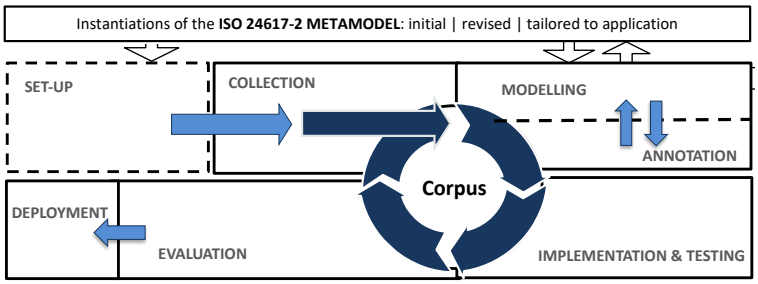

Figure 4.1: Continuous dialogue corpus creation (D3C).

tasks, recording setting (equipment and environment) and description of data collection process;

2. **Collect**: record, encode and store human-human dialogue primary[2] data for the specified scenario;
3. **Model**: extend the standard ISO 24617-2 metamodel using the annotations produced. In particular, include SemAF concepts from the application domain;
4. **Annotate**: apply standard and domain-specific annotation scheme(-s) to classify a particular set of entities and their properties;
5. Implement & Test: build (train) and test dialogue system components and resulting dialogue models, utilising annotations produced in the previous steps; optionally experiment with tuning components of the dialogue system;
6. **Evaluate**: perform objective (system performance) and user-based (user perception) evaluation with the integrated dialogue system prototype in the laboratory and close to operational environments; log evaluation sessions and analyse results;
7. **Deploy** (*optional after each iteration*): write to the corpus, document and prepare to be released including signals, primary data, annotations and corpus manual with schemes, guidelines and format specifications;
8. Repeat steps 1-7 for the full cycle for a refined set-up, or steps 3-6 to re-train system modules based on data obtained in user-based evaluation sessions.

The proposed methodology is in the line with principles of semantic annotation defined in the ISO standard 24617-6 which characterises the ISO semantic annotation framework [ISO, 2016]. The standard includes the CASCADES (**C**onceptual analysis, **A**bstract syntax, **S**emantics, and **C**oncrete syntax for **A**nnotation language **DES**ign) annotation schemes design model [Bunt, 2015]. The model enables a systematic (re-)design process: from conceptual ('metamodel') and semantic choices ('abstract' syntax) to more superficial decisions such as the choice of particular XML attributes and values ('concrete' syntax), see Figure

[2]Data observed or collected directly from first-hand experience such as representation of written (e.g. text), spoken (e.g. orthographic transcriptions of audio) and multimodal (e.g. images or videos) behaviour. Typically, primary data objects are represented by "locations" in an electronic file, e.g. the span of characters comprising a sentence or word, or a point at which a given temporal event begins or ends. More complex data objects may consist of a list or set of contiguous or non-contiguous locations in primary data, see [Ide and Romary, 2004]

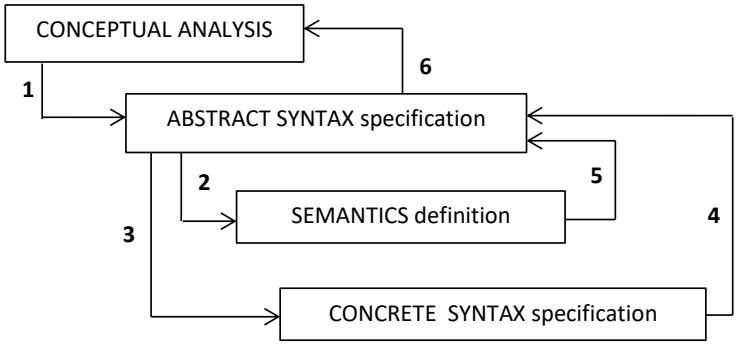

Figure 4.2: CASCADE design method, Bunt (2005).

4.2. The method can be used to design a new annotation scheme or provides support for improving an existing annotation scheme through feedback loops.

The CASCADES is integrated with the MATTER method [Pustejovsky et al., 2017] for annotation and data modelling, conceptualised as the Model, Annotate, Train, Test, Evaluate and Revise cycle which inspired the methodology presented here.

4.2 The ISO 24617-2 data model

Standard well-specified and evaluated data models are the key enablers for corpus and system development. They are a prerequisite for the corpus to be of good quality, provides ways to systematically incorporate extensions, and ensure interoperability, enabling sharing, merging and comparison with other resources. Data models, formalized descriptions of data objects and relations between them, are designed to capture the structure and relations in diverse types of data and annotations. Well-specified standard resource formats and processes facilitate the exchange of information between different processing modules. Mappings between documents containing primary data[3] and the data model are operationalized via schema-based data-binding processes (Ide and Romary, 2004).

4.2.1 Basic concepts

As previously discussed, dialogue acts are key notions in the description of the meaning of dialogue utterances and are central to the ISO 24617-2 standard metamodel. The ISO standard defines a dialogue act as

(4) *communicative activity of a participant in dialogue interpreted as having a certain communicative function and semantic content, and possibly also having certain functional dependence relations, rhetorical relations and feedback dependence relations.*

[3]Data observed or collected directly from first-hand experience such as representation of written (e.g. text), spoken (e.g. speech transcriptions) and multimodal (e.g. images, videos) behaviour.

A dialogue act has at least two participants: (1) an agent whose communicative behaviour is interpreted, usually called the *speaker*, or *sender*; and (2) a participant to whom he is speaking and whose information state he wants to influence, called the *addressee* (also called "hearer" or "recipient"). There may of course be more than one addressee, e.g. in debate situation there is 'audience' and 'opponents' who are addressees of the 'debater' communicative actions.

Besides sender and addressee(s), there may be various types of *side-participants* who witness a dialogue without participating in it. The presence of side-participants may influence the communicative behaviour of the participants, if these are aware of their presence, as in a television interview or a talk show. Clark (1996) distinguishes between 'side-participants', 'bystanders', and 'overhearers', depending on the role that they play in the communicative situation.

A dialogue act has two main component: communicative function and semantic content. A *communicative function* specifies the way semantic content is to be used by the addressee to update his context model when he understands the corresponding aspect of the meaning of a dialogue utterance. *Semantic content* indicates what the behaviour is about: which objects, events, situations, relations, properties, etc. Annotation of semantic content is concerned with annotating different natural language phenomena like events, named entities, semantic roles, semantic relations, etc. Semantic content of a certain type (called *dimension*) is "an abstract characterisation of the content of an utterance" (Allen and Core, 1997) and may address information about a certain *Task*; the processing of utterances by the speaker (*Auto-feedback*) or by the addressee (*Allo-feedback*); the management of difficulties in the speaker's contributions (*Own-Communication Management*) or that of the addressee (*Partner Communication Management*); the speaker's need for time to continue the dialogue (*Time Management*); the allocation of the speaker role (*Turn Management*); the structuring of the dialogue (*Dialogue Structuring*); and the management of social obligations (*Social Obligations Management*), see ISO, 2012. The ISO dimension set can be extended. For specific purposes or domains, new dimensions may be added. One property that a potential additional dimension should satisfy is that it should be orthogonal to the already existing dimensions, in order to avoid redundancy and ambiguity in annotation; for orthogonality tests see (Petukhova, 2011). A dimension and the corresponding set of dimension-specific communicative functions may be freely leftout and this has no influence on the remaining dimensions due to their orthogonality.

In dialogues, dialogue acts are not produced in isolation, but various relations exists between them. The meaning of a responsive dialogue act such as Answer, Accept Request or Agreement typically depends on the meaning of a previous dialogue act (or dialogue acts) such as Question or Request respectively. To represent such dependencies a *functional dependence relation* is defined in (ISO, 2012) as:

(5) *A functional dependence relation exists between a dialogue act DA_1 and one or more previous dialogue acts $\{DA_2, ..., DA_N\}$ iff the meaning of DA_1 depends on the meaning of $\{DA_2, ..., DA_N\}$ due to the responsive character of DA_1.*

Responsive dialogue acts of another type provide or elicit information about the (per-

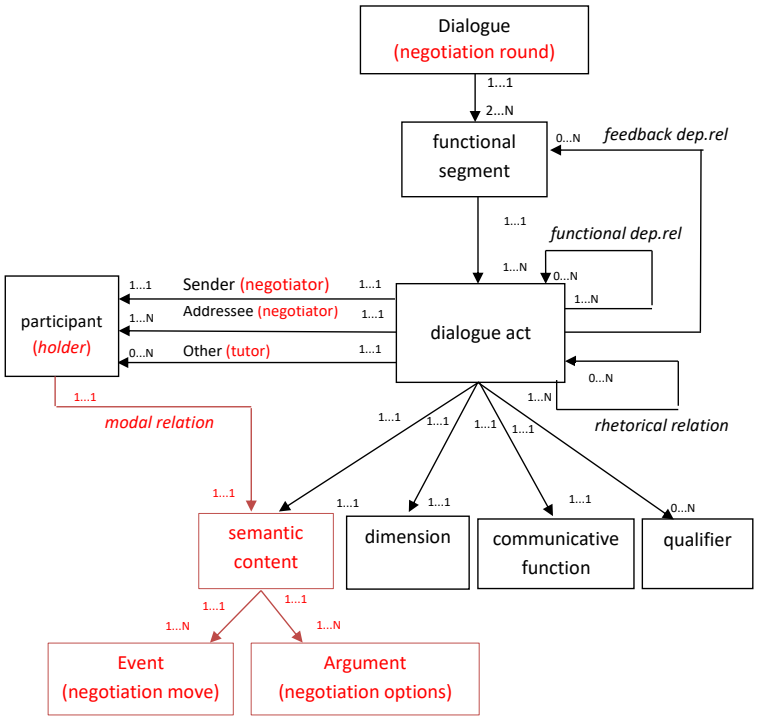

Figure 4.3: ISO 24617-2 metamodel for dialogue act annotation. Domain-specific extensions marked red.

ceived) success in processing an earlier segment (or segments) of communicative behaviour. Such a relation is called a *feedback dependence relation*. This type of relation has been defined in ISO standard 24617-2 as follows:

(6) *A feedback dependence relation is a relation between a feedback act and the stretch of communicative behaviour whose processing the act provides or elicits information about.*

Feedback acts refer explicitly or implicitly to the stretch of dialogue behaviour that they provide or elicit information about. This stretch of dialogue behaviour forms part of its semantic content. It has been observed recently that ISO 24617-2 has certain shortcomings with respect to dependence relation definitions, see Bunt et al., 2017. It has been concluded that a dialogue act can have a functional or a feedback dependence relation, but not both. This would make it possible to drop the terminological distinction and just speak of 'dependence relation'.

Rhetorical relations may be annotated to indicate, for example, that one dialogue act explains the performance of another dialogue act such as explanation, justification, cause, etc. A set of rhetorical relations is defined in ISO 24617-8 - Language Resource Management - Semantic Annotation Framework - Part 8: Semantic relations in discourse, Core annotation scheme (ISO DR-core) Bunt and Prasad, 2016. For example[4]:

[4]From UK Youth Parliament Debates, see Petukhova et al., 2015

(7) $D2_{30}$: Essentially we are experiencing a tragic loss of childhood [*Inform*]
$D2_{31}$: a walk down the high street reveals a depressing trend towards essentially adult's designs [*Inform Evidence* $D2_{30}$]
$D2_{32}$: children's pencil cases bearing playboy symbols [*Inform Evidence* $D2_{30}$*; Inform Motivate* $D2_{31}$]
$D2_{33}$: our children being sexualized too young [*Inform Result* $D2_{30}$*,* $D2_{31}$,$D2_{32}$;*Cause* $D2_{34}$]
$D2_{34}$: we must aim to protect this short-lived innocence [*Inform Result* $D2_{33}$]
$D2_{35}$: SRE is simply inappropriate within a primary curriculum [*Inform Conclude* $D2_{30}$ - $D2_{34}$]

Speaker's intentions may be rather complex, vague and ambiguous. They may also be emotionally qualified expressing particular attitudes towards their communicative partners, third parties and message content. To capture nuances in meaning description which concerns certainty, conditionality and sentiment, *qualifiers* are introduced in ISO 24617-2.

To sum up, in the characterization of the notion of a dialogue act and its realization, as given so far, the following key elements occur:

- sender (or 'speaker')
- addressee(s)
- participants in other roles (such as overhearers)
- functional segment
- dialogue act
- communicative function
- communicative function qualifier
- semantic content type
- dependence relations
- rhetorical relations between dialogue acts

Formalized descriptions of defined semantic objects and relations between them are captured in the ISO 24617-2 data model (or 'metamodel', see Figure 4.3) which represents the fundamental upper-level concepts that are involved in dialogue act annotation. Thus, a dialogue consists of two or more functional segments. Each segment is related to one or more dialogue acts, reflecting the possible multifunctionality of functional segments. Each dialogue act has exactly one sender, one or more addressees, and possibly other participants. It has a semantic content of a certain type ('dimension'), and a communicative function, which may have any number of qualifiers. Dialogue acts are possibly related to other dialogue acts through functional dependence and rhetorical relations, and to functional segments through feedback dependence relations.

The ISO 24617-2 model is extended as shown in Figure 4.3 where proposed extensions are marked red. It mainly concerned the specification of the semantic content of a dialogue act which is domain-specific. Semantic content can be specified in terms of predicate-argument structures, named entities, semantic roles, etc., applying other available standards

of the ISO Semantic Annotation Framework. An example of domain-specific semantics is provided for negotiation dialogues in terms of negotiation events such as offer, counter-offer, concession, etc., and their arguments in Petukhova et al. (2017). Additionally, the classified modality related to the speaker's preferences, priorities, needs and abilities is defined in Lapina and Petukhova (2017).

4.2.2 ISO Dialogue Act Markup Language

The Dialogue Markup Language (DiAML) [ISO, 2012] is used as the representation and exchange format in dialogue corpus and system development; DiAML is also used for communication among all system modules, and for representing intermediate and end results.

According to ISO 24617-2, the representation of a dialogue act annotation with the Dialogue Act Markup Language (DiAML) makes use of the XML element `<dialogueAct>`. This element has the following attributes:

- `@target`, whose value is a functional segment identified at the second level;
- `@sender`, `@addressee`, `@otherParticipant`;
- `@communicativeFunction`, `@dimension`;
- `@certainty`, `@conditionality`, and `@sentiment` qualifiers;
- `@dependenceRelation` which has `<dialogueAct>` elements and `<functionalSegments>` as values.

Rhetorical relations among dialogue acts are represented by means of `<rhetoLink>` elements. All these types are defined in `diaml` namespace in the defined `DiAML_Types.xsd` scheme. Elements in the `DiAML_Containers.xsd` specified without a namespace. This allows for unifying DiAML_Containers with other (not-`diaml`) schemas and that is, usually, for purposes of specifying XML elements that express domain-dependent semantics. A sensible coverage of the domain-dependent semantics seems to be achieved by declaring elements of the semantic content of dialogue acts. In turn, such an approach allows for automated data validation and automated generation of programming code that defines object classes used for both communicating and processing of data. DiAML annotations were extended with a semantic content, also shown by Bunt et al. (2017). Consider the following ISO DiAML representation as an example using an `<aSemantics>` element:

```
<dialogueAct xml:id="dap1" sender="#p1"
         addressee="#p2" dimension="task"
         communicativeFunction="inform"
         target="#fsp1">
  <aSemantics>
    <event xml:id="e1" type="offer">
      <arg>10_percent</arg>
      <modalLink holder="#p1" target="#e1"
                 modalRel="preference"/>
    </event>
  </aSemantics>
</dialogueAct>
```

The <event> element, which specifies information about the semantic content of a dialogue act can be defined based on the ISO annotation schemes for time and events (ISO 24617-1), for semantic roles (ISO 24617-4), and for spatial information (ISO 24617-7), and that has also been proposed for the annotation of modality (Lapina and Petukhova, 2017) and quantification (Bunt, 2017). This opens the possibility to specify quite detailed information about the semantic content of dialogue acts, including domain-specific semantics as shown by Petukhova et al. (2017a) for negotiations. The <negotiationSemantics> element has been defined to represent the semantic content of a dialogue act. A shallow negotiation semantics is defined in terms of <negotiationMove> with attributes defined for different types of such moves. For example:

```
<dialogueAct xml:id="dap1TSK38" sender="#p1"
             addressee="#p2" dimension="task"
             communicativeFunction="inform"
             target="#fsp1TSKCV38">
  <negotiationSemantics>
    <negotiationMove type="counterOffer"/>
  </negotiationSemantics>
  <rhetoricalLink rhetoAntecedent="#dap2TSK37"
                  rhetoRel="substitution"/>
</dialogueAct>
```

Additionally, dependent on annotation goals, approach, granularity and type of semantic processing, <negotiationSemantics> elements can be extended with elements, based on <Arg> type, for negotiated issues, values and logical operators between arguments.

Domain-specific semantics for the debate domain, another domain considered in this thesis, is concerned with 'for' and 'against' arguments featuring a certain topic. This information can be also plugged into DiAML using the <debateSemantics> element. For example:

```
<dialogueAct xml:id="da1" sender="#p1"
             addressee="#p2" dimension="task"
             communicativeFunction="inform"
             target="#fs38" qualifier="certain">
  <debateSemantics>
    <argument type="for" topic="tax_increase" />
  </debateSemantics>
</dialogueAct>
```

4.3 Use Case: interactive training of metacognitive skills

4.3.1 Interactive learning and tutoring

Cognitive Tutoring Systems aim to support the development of metacognitive skills. Examples of such systems are described in Bunt and Conati (2003), Azevedo et al. (2002),

Gama (2004), Aleven et al. (2006) and Baker et al. (2006). These systems rely on artificial intelligence and cognitive science as a theoretical basis for analysing how people learn (Roll. et a., 2007).

Research by Chi et al. (2001) revealed that the interactivity of human tutoring drives its effectiveness. *Interactive learning* is a modern pedagogical approach that has evolved out of the hyper-growth in the use of digital technology and virtual communication. Interactive learning is a promising and powerful way to develop metacognitive skills. In this study, the interactivity of a tutoring system is achieved through the use of multimodal dialogue. While many intelligent tutoring dialogue systems have been developed in the past (Litman and Silliman, 2004; Riedl and Stern, 2006; Core et al., 2014; Moore et al., 2005; Paiva et al., 2004), to the best of our knowledge no existing cognitive tutoring system makes use of natural spoken and multimodal dialogue.

Metacognitive skills are domain-independent and should be applicable in any learning domain and in a variety of different learning environments, but despite their transversal nature, metacognitive skills training can only be practiced within certain domains and activity types. Some systems have been developed successfully for domains such as mathematics, physics, geometry, biology and computer programming (MetaTutor, Azevedo et al., 2009; Rus et al., 2009: Harley et al., 2013). For debate and negotiation, metacognition has been empirically proven to be important since it significantly improves decision-making processes (Aquilar and Galluccio, 2007).

4.3.2 Debate training

The Debate Coach Agent (DCA) is designed to be integrated as a part of the Virtual Debate Coach application - an intelligent tutoring system used by young parliamentarians to train their debate skills. A debate is a communication process in which participants argue for or against a certain position proposed for the dispute. Whereas the argumentative elements of debating have received ample attention as a means to enhance learning [D'Souza, 2013], learning relevant aspects of debating are understudied.

Few existing argumentation training systems work for spoken discourse. For example, Ashley et al. [Ashley et al., 2007] use transcripts of arguments produced in the US Supreme Court as a basis for training hypothetical reasoning drawing the similarities with the legal case in question. The trainee is shown an argument transcript and asked to build an argumentation structure graph following Toulmin's scheme (see below). The system detects trainee's contextual and structural weaknesses, and provides feedback. Trainees do not formulate their own arguments but use pre-defined phrases, or are offered the option to substitute special legal formulations with semantically equivalent ones.

There are web-based argumentation training systems available, e.g. DebateGraph[5] and TruthMapping[6]. The former aims at providing a platform to prevent opinion manipulation, marking inconsistent arguments in online forum discussions. The system represents arguments as graphs including unsupported premises and giving the user the possibility to rebut

[5] `http://debategraph.org/`
[6] `https://www.truthmapping.com/`

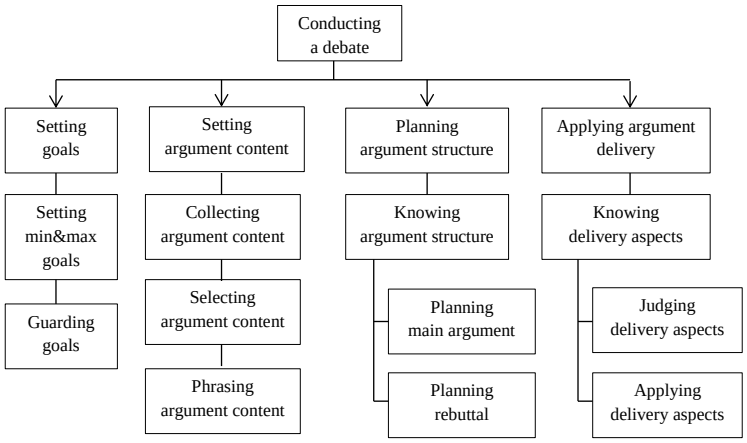

Figure 4.4: Hierarchy of skills involved in debating. Adapted from Van Rosmalen et al., 2015

or support arguments. TruthMapping facilitates collaborative learning through argumentation. Arguments are also represented as graphs, different standpoints and their evidences are visualised to the learners encouraging them to address those.

In our scenario, debaters exchange 'natural' arguments, i.e. they are not constrained in the use of communicative means, and they also may exploit 'extra-rational' characteristics of their audience, taking into account emotions and affective factors. It is important, therefore, not only to understand the underlying structure of natural arguments explaining certain regularities but also to evaluate means and strategies used by debaters to deliver convincing performance. The training of debate skills typically involves ad-hoc face-to-face classroom debates.

In current educational design practice there is a growing interest in using whole-tasks models. Whole-tasks models aim to assist students in integrating knowledge, skills and attitudes into coherent wholes, to facilitate transfer of learning. As part of this they take into account how to balance the load of the learner, make the tasks sufficiently challenging and how to give feedback. Characteristics of a 'skilled professional debate performance' are defined in terms of coaching goals related to (1) argument organisation, (2) argument content, and (3) argument delivery.

Conducting a debate is a complex task. The skill to be mastered is in brief "convincingly present, argue and respond about a current hot issue". For this, a trainee needs to have knowledge and skills about both argument content and structure aspects (e.g. what to present, how to use and structure their arguments, how to rebut, what and how to close the argument) and delivery aspects (e.g. how to use their voice e.g. pitch, speed or volume, body etc). On top of this, continuously, the trainee has to be aware of the effects of their debating inputs and guard their goals by monitoring the level of agreement, both content wise but also how they and their opponents respond and reflect and adapt accordingly when necessary. The skills required to perform this task adequately can be seen as formed arount four foci (see Figure 4.4 for the 'conducting debate' skills hierarchy proposed in [Van Rosmalen et al., 2015]):

- Setting goals: set and guard the desired target with regard to the aim of the dialogue (e.g. pass a proposal with as little changes as possible) and the ability of the learner

to anticipate on an opponent and adapt accordingly to achieve the goal;

- Setting argument content: search, select and phrase the relevant content;
- Planning argument organisation: organise content, arguments, counter-arguments and objections;
- Applying argument delivery: present the content taking into account delivery aspects.

Argument and argumentation

An argument is defined as consisting of a statement that can be supported by evidence. A statement or *claim* is an assertion that deserves attention. There may be a conclusion which presents some kind of result, which can be derived from certain evidence (*premises*).

Previous work in argumentation theory and artificial intelligence was largely based on designing and applying argumentation schemes, see Toulmin (1958), Walton (1996) and Freeman (2011). Toulmin (1958) proposed a scheme with six functional roles to describe the structure of an argument. Based on evidence (*data*) and a generalisation (*warrant*), which is possibly implicit and defeasible, a *conclusion* is derived. The conclusion can be *qualified*, e.g. by strengthening the inferential link between data and conclusion. A *rebuttal* specifies exceptional conditions that undermine this inference. A warrant can be supported by *backing*, e.g. reason, justification or motivation. Figure 4.5 depicts Toulmin's argumentation scheme.

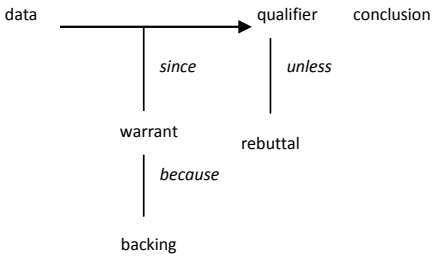

Figure 4.5: Argumentation diagram of Toulmin (1958).

Recently, argumentation mining techniques have been applied to natural arguments analysis, see the survey in [Peldszus and Stede, 2013]. Independent of the approach, most researchers seem to agree on the theoretical skeleton of logical and pragmatic aspects - the connection between subject and predicate on a logical propositional level and the inter-propositional relations on the pragmatic level. Translating Toulmin's general argumentation scheme into a structure of debate arguments, we have premises for a claim (main statement, **A***rgument*) that can be of **R***eason* and **E***vidence* types, and a claim that may be summarised or re-stated in a conclusion, often referred as an ARE structuring technique, see Figure 4.6.

Another commonly used technique to support a claim with evidence is called *chunking* [Johnson, 2009]. Here, debaters generalise from a claim (*chunking up*), provide a specific example (*chunking down*) or draw analogies (*chunking sideways*).

Debaters are trained to follow rules imposed by the above mentioned structures, respect domain conventions and best practices.[7]

[7] See the debate competition guidelines of the English Speaking Union `http://www.esu.org/`

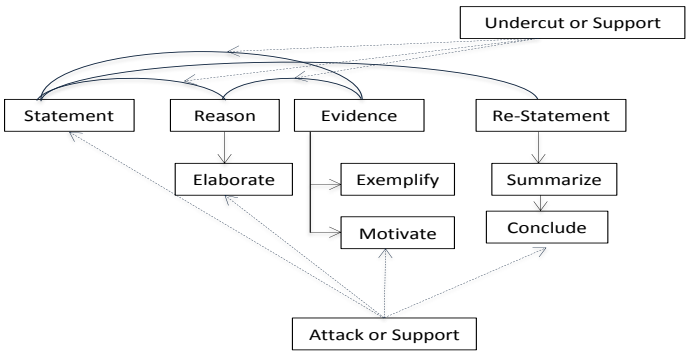

Figure 4.6: Analysis of argument structure. Adapted from Petukhova et al., 2015

Delivering convincing debate performance

Debates, in particular political debates, constitute a large portion of public speeches. Skilled professional debaters give the impression that they truly believe what they say, know how to catch and keep the attention of the audience, express authority, confidence, respect and friendliness. People generally associate certain speech, personality and interaction features with what they think is a 'good public speaker' [Strangert and Deschamps, 2006]. Debaters make a number of choices from a wide range of rhetorical, lexical, syntactic, pragmatic and prosodic devices to deliver strong persuasive speech. They often use *intensifiers*, i.e. individual words or phrases that are syntactically, tonally or rhythmically marked, *parallelisms* (word or phrase repetitions for information density reduction and emphasis, e.g. well-known 'Lists of Three' [Beard, 2002]), and *meta-discursive acts*[8] to relate speaker to audience, to maintain topic-comment structure, etc. [Nir, 1988, Beard, 2002, Touati, 2009]. Prosodic and acoustic strategies in speech may be decisive in conveying an opinion in a political debate [Braga and Marques, 2004]. Clear articulation, sufficient voice volume level, and well adjusted tempo are strongly associated with professional public speaking. Pitch range, voice and speaking rate variations are perceived as expressions of enthusiasm, engagement, commitment and charisma, see also [Rosenberg and Hirschberg, 2009]. Mispronounced words, frequent hesitations, restarts and self-corrections negatively influence the perceived speaker confidence and may jeopardise speaker credibility [Tuppen, 1974].

Effects of audio-visual prosody on the perception of information status related to focus and prominence have been also studied. For example, investigating *visual beats* it has been concluded that if observers see a visual beat they perceive a corresponding phrase as more prominent [Krahmer and Swerts, 2007]. We may expect that prosodically prominent phrases when accompanied by gestures will intensify the assertiveness and persuasion effect of the debate arguments. Good debaters that score high on expression and delivery demonstrate a clear awareness of rhetoric and attempt to engage an audience. They make use of direct eye contact, body language and emotive language.[9] Persuasive debate performance

[8] Crismore et al. (1993) define metadiscourse as "linguistic material in texts, written or spoken, which does not add anything to the propositional content but that is intended to help the listener or reader organise, interpret and evaluate the information given", e.g. Shifting Topic, Marking Asides, etc.

[9] See `http://www.esu.org/\_data/assets/pdf\_file/0011/16202/ESU\`

may be linked to dominance. Crossing the arms, stemming the hands on the hip or touching one's neck most effectively influence dominance perception [Straßmann et al., 2016].

4.3.3 Multi-Issue Bargaining

Three main types of negotiations can be distinguished: *distributive*, *joint problem-solving* and *integrative*[10]. Distributive negotiation means that any gain of one party is made at the expense of the other and vice versa; any agreement divides a fixed pie of value between the parties, see e.g. [Walton and McKersie, 1965]. The goal of joint problem-solving negotiations is, by contrast, to work together on an equitable and reasonable solution: negotiators will listen more and discuss the situation longer before exploring options and finally proposing solutions. The relationship is important for joint problem solving, mostly in that it helps trust and working together on a solution, see [Beach and Connolly, 2005]. In integrative bargaining, parties bargaining over several goods and attributes search for an integrative potential (interest-based bargaining or win-win bargaining, see Fisher an Ury, 1981). This increases the opportunities for cooperative strategies that rely on maximising the total value of the negotiated agreement (enlarging the pie) in addition to maximising one's own value at the expense of the partner (dividing the pie).

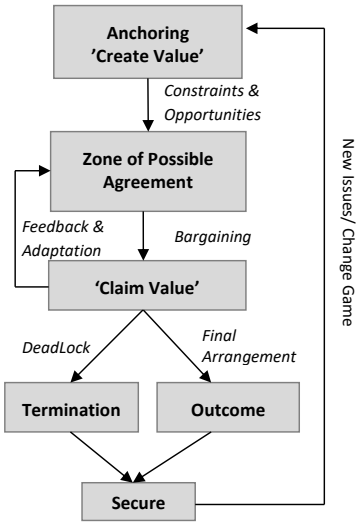

Figure 4.7: Negotiation phases associated with negotiation structure, based on [Watkins, 2003, Sebenius, 2007].

The different types of negotiation are manifesting mainly in how parties create and claim values. Negotiation starts with the **Anchoring** phase, in which participants introduce negotiation issues and options. They also obtain and provide information about preferences, establishing jointly possible values contributing to the **Zone of Possible Agreement** (ZOPA, Sebenius, 2007). Participants may bring up early (tentative) offers, typically in the form of suggestions, and refer to the least desirable events - 'Create Value'. The actual bargaining occurs in the **'Claim Value'** phase, potentially leading to adaptation, adjustment or cancelling the originally established ZOPA actions. Patterns of concessions, threats, warnings, and early tentative commitments are observed here. Distributive negotiations are more 'claiming values', while joint problem-solving negotiations are more 'value creating' interactions, and integrative negotiations are a mix of 'creating and claiming values' negotiations (Watkins, 2003a). In distributive negotiations, the size of the possible agreement range is mostly determined by the 'bottom lines' of

_Debate_Challenge_2017_v2.pdf

[10] A fourth type of negotiations is *bad faith*, where parties only pretend to negotiate, but actually have no intention to compromise. Such negotiations often take place in political context, see [Cox, 1958]

the opposite parties, which are formed by their respective *best alternatives to a negotiated agreement* (BATNA), see [Fisher and Ury, 1981]. In integrative bargaining, this range/zone is mainly determined by the number of possible Pareto optimal outcomes. Pareto optimality reflects a state of affairs when there is no alternative state that would make any partner better off without making anyone worse off.

After establishing the ZOPA, negotiators may still cancel previously made agreements and negotiations might be terminated. **Negotiation Outcome** is the phase associated with the "walk-away" positions for each partner. Finally, negotiators can move to the **Secure** phase summing up, restating negotiated agreements or termination outcomes. At this stage, strong commitments are expressed, and weak beliefs concerning previously made commitments and agreements are strengthened. Participants take decisions to move with another issue or re-start the discussion. Figure 4.7 depicts the general negotiation structure as described in [Watkins, 2003] and [Sebenius, 2007], and observed in our data described in the next section.

The negotiation outcome depends on the setting, but also on the agenda and the strategy used by each partner (Tinsley et al., 2002). The most common strategy of novice negotiators observed is issue-by-issue bargaining (see data collection below). Parties may start with what they think are the 'toughest' issues, where they expect the most sharply conflicting preferences and goals, or they may start to discuss the 'easiest', most compatible options. Sometimes, however, negotiators bring all their preferences on the table from the very beginning. This increases the chance to reach a Pareto efficient outcome, since a participant can explore the negotiation space more effectively, being able to reason about each others' goals, see e.g. [Stevens et al., 2016b]. Defensive behaviour, i.e. not revealing preferences, but also being misleading or deceptive, i.e. not revealing true preferences, results in missed opportunities for value creation, see e.g. [Watkins, 2003, Lax and Sebenius, 1992]. It has been also observed that as a rule it is easier for a negotiator to bargain down, i.e. to start with his highest preference and if this is not accepted by the partner, go down and discuss sub-optimal options, than it is to bargain in, i.e. to reveal his minimum goal and go up, offering preferences that are not necessarily shared by the partner.

4.4 Set-up and scenarios

The design of any system requires a clear understanding of the users, their goals and the usage situation. This helps to determine the system's functionality, reduces design mistakes and often provides good inspiration and orients. The data collection set-up includes first of all the specification of the intended users and system requirements. A users analysis is conducted to define key user groups (age, gender, cultural and educational backgrounds, etc.) and identify their interest areas, known attitudes, values and priorities. Context of use, settings and users' needs have a direct impact on the role the system will play in an interactive situation, and therefore on the system functionality. Apart from the pure communicative tasks that a dialogue system has, to understand and to react to users' intentions, a dialogue system has tasks dependent on the application domain in relation to the role(-s) it plays, e.g.

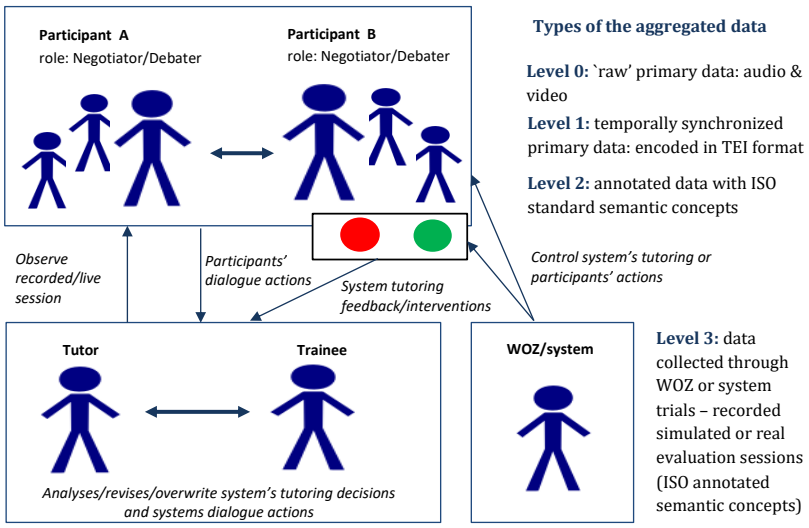

Figure 4.8: Example of the system and data collection set-up.

as an assistant, adviser or mediator, as a passive observer, as a tutor or as a coach. Users, context and system requirements are used not only to make important design decisions but also to define appropriate verification and evaluation strategies. The evaluation tasks should be representative for most users such that results can be generalised beyond the specific sample.

The 24617-2 ISO data model forms the basis for a domain-specific data collection set-up specifying the type of interaction, participants roles, tasks and actions performed. For example, in our negotiation training scenario, we have a negotiation *session* consisting of one or multiple training *rounds* featuring different goals assigned to trainees by a *Tutor*. Tutors (humans or simulated agents) attend the session and provide feedback to *Trainees* performing a negotiation or debate task. Tutoring interventions are expected to inform trainees of mistakes, propose corrections, provide instructions, initiate 'try again' rounds, or highlight trainees' successes. This involves immediate real-time 'in-action' and summative 'about-action' feedback (Schön, 1983). The task of trainees as *Negotiators* is to propose offers and react to offers of the partner, and as *Debaters* to propose arguments in favor or against a certain issue (see below). An extended ISO 24617-2 metamodel (see concepts marked red in Figure 4.3) underlies all system and corpus development. A general framework for data collection is set up as shown in Figure 4.8. We specify participant roles and tasks, as well as data types collected at each recording, as well as processing and evaluation stages including simulated and real dialogue system behaviour in the role of tutor and participant.

The technical set-up specifies recording conditions, equipment, instructions for technical personnel, as well as details on type and granularity of data that should be recorded, and how and where it should be stored, see [Haider et al., 2017].

Issue under debate	Trainees minimum goals to achieve		
	Proponent	Opponent (conservative)	Opponent (liberal)
Smoking ban scope	Not all public places should be affected, allow smoking in bar and restaurants and open air areas like outside buildings, parks and beaches	Forbid smoking inside all public spaces, special smoking areas outside buildings	Allow smoking in special areas in bars and restaurants, open air places also need smoking areas
Tobacco prices	Tobacco price already high, increase no more than 2% a year	Tobacco prices are low, increase by 10% a year	Tobacco prices are still too attractive, increase by 5% a year
Access to tobacco	Tobacco sold in supermarkets, specialised licensed tobacco shops, in bars and restaurants, and vending machines on street with secured buyer's age control	Tobacco should be sold only in special licensed tobacco shops	Tobacco sold in supermarkets but hidden in special containers, prohibited to sell around schools (5km distance) and not available in bar or street vending machines
State control	No police control but municipal and administrative control, no penalties but warnings for the 1st time, repeated disobedience may be punished with penalties	Strong police presence in public places and penalties without warnings	No police control, municipal and administrative control, 1st time disobedience gets warning; second time penalties
Anti-smoking campaign	on TV (state channels 20 min broadcasting time a week); posters in every public place; 'educated' slogans on cigarettes; big newspapers 5 lines a week on the first 2-3 pages	on TV (all channels 30 min broadcasting time a week + one documentary a month); posters in every public place; slogans and scaring images on cigarettes; big newspapers 10 lines a week on the bottom of the front page	on TV (state channels 20 min broadcasting time a week); posters in every public place; 'educated' slogans on cigarettes; big newspapers 10 lines a week on the first 2-3 pages

Table 4.1: Example of participants' minimal goals in one debate round.

4.4.1 Debate scenario

The specific setting considered for the data collection involves a debate scenario about anti-smoking legislation in Greece. The initial proposal for a smoking ban is supported by the proposing (governmental) party. The goal of the proposer is to get a majority vote while agreeing on as few amendments as possible.

Our core data collection activity involved debate *trainees*, school children aged 14-15 years who have been exposed to little debate training. A session involved a pair of participants: one assigned the role of proposer, the other the role of either liberal or conservative opponent. Each participant was given a set of minimal goals concerning: (1) a total ban on smoking in public spaces; (1) limiting youth access to tobacco products; (3) improving the effectiveness of anti-smoking campaign; (4) state control and reinforcement policy; (5) and raising prices on tobacco products. Participants were not allowed to disclose their goals to the other parties prior to the interaction. Three human tutors evaluated debate performance. Table 4.1 provides an example of minimal goals that trainees playing different roles should achieve in one debate round.

The collected data consists of 12 sessions with a duration of 2.5 hours, comprising 400 arguments (Argumentative Discourse Units, ADUs[11]) from 6 different bilingual English/-Greek speakers.

4.4.2 Negotiation scenario

For adequate modelling of human-like multi-issue bargaining behaviour, a systematic analysis of collected and semantically annotated human-human dialogue data was performed. The specific setting considered in this study involved a real-life scenario about anti-smoking legislation in the city of Athens passed in 2015-2016. After a new law was enacted, many cases of civil disobedience were reported. Different stakeholders came together to (re-)negotiate and improve the legislation. The main negotiation partner was the Department of Public Affairs of the City Council who negotiates with representatives of small businesses, police, insurance companies, and others.

The anti-smoking regulations were concerned with four main *issues*: (1) smoke-free public areas (scope); (2) tobacco tax increase (taxation); (3) anti-smoking program promotion (campaign); and (4) enforcement policy and police involvement (enforcement), see Figure 4.9. Each of these issues involves four to five most important negotiation *values* with preferences representing negotiation positions, i.e. preference profiles. Nine cases with different preference profiles were designed. The strength of preferences was communicated to the negotiators through colours. Brighter orange colours indicated increasingly negative options; brighter blue colours increasingly positive options.

In the data collection experiments, each participant received the background story and a preference profile. Their task was to negotiate an agreement which assigns exactly one value to each issue, exchanging and eliciting offers concerning $\langle ISSUE; VALUE \rangle$ options.

[11] For more details on segmentation and annotation performed, we refer to [Petukhova et al., 2016a].

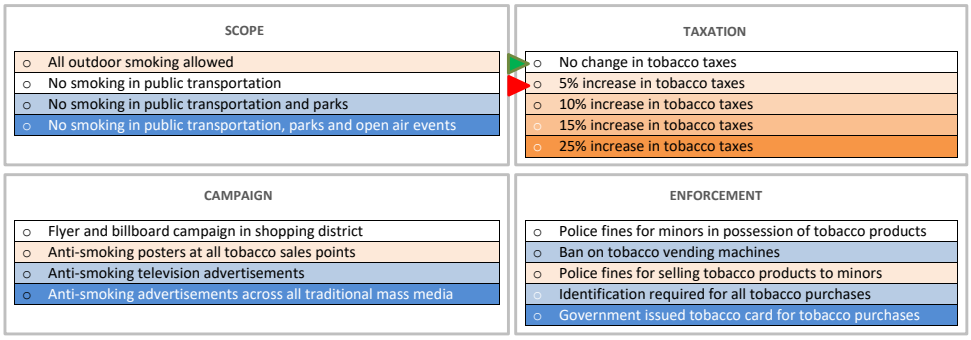

Figure 4.9: Preference card: example of values in four negotiated issues presented in colours: brighter orange colours indicated increasingly negative options and brighter blue colours increasingly positive options. When incorporated into the graphical interface, partners' offers visualised with red arrow (system) and green one (user).

Participants were randomly assigned their roles. They were not allowed to show their preference cards to each other. No further rules on the negotiation process, order of discussion of issues, or time constraints were imposed. They were allowed to withdraw or re-negotiate previously made agreements within a session, or terminate a negotiation.

16 subjects (young professionals aged between 19 and 25 years) participated in the experiments. The resulting data collection consists of 50 dialogues of a total duration of about 8 hours, comprising approximately 4.000 speaking turns (about 22.000 tokens).

4.5 Collection and processing

In multimodal dialogue applications, speech is the main modality. Speech recordings should be of sufficient quality to be used for further processing. Our experience is that recorded 96KHz/24bits audio signals allow a very good tracking of prosodic variations and can be down-sampled to train an Automatic Speech Recognition (ASR) system.

Speech was captured by two audio Tascam Dr-40 recorders and two Sennheiser headsets, and saved in MS WAV format[12]. Speech files are of two types: (1) full dialogue session recorded per speaker, and (2) cut audio files per speaker and roughly per turn (after speaker diarization). Speaker diarization has been partly carried out manually using the Audacity tool[13] and partly automatically using the LIUM tool [Rouvier et al., 2013]. The speech signal files contain timestamps - start and end time - and additional comments on acoustic and temporal conditions (noise, long silences, etc.) in the file name. For example, `08.22-08.30.n.wav` is the segment which started at 8 minutes and 22 seconds and finished at 8 minutes and 30 seconds during the recording session; and it contains some noise indicated by "n". The speech of a dialogue participant was transcribed semi-automatically

[12]The recordings were performed in the following setting: sample rate (48KHz), sample size (16-bit), sample format (linear PCM) with stereo channel which was later converted to mono .

[13]`http://www.audacityteam.org/`

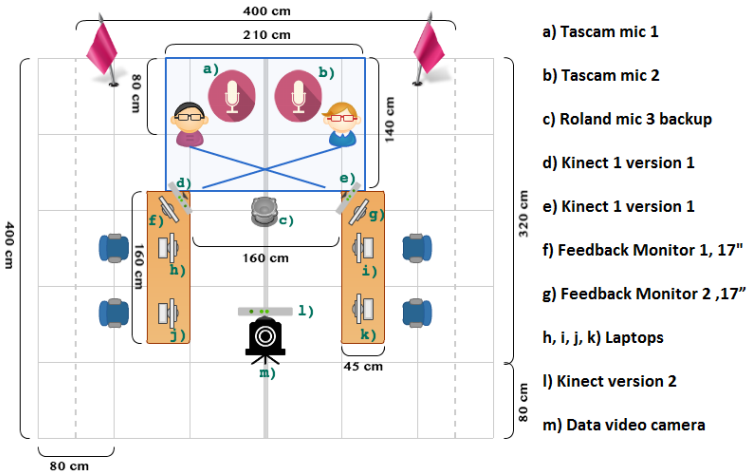

Figure 4.10: Recording set up for training sessions, adapted from Haider et al. (2017).

by (1) running the ASR system and (2) correcting transcriptions manually. All transcription were stored per participant and dialogue session in TEI compliant format [ISO, 2006].

Training sessions were recorded in a quiet room under special lighting conditions, ensuring that there were no windows behind the participants and that the participants' faces were not in shadow. Two Kinect $V1$ sensors, each facing one participant as much as possible, were placed at a distance of 1.5-2m to the participants. A Kinect $V2$ sensor was also used to track both participants. Body and face tracking data were stored in an XML format containing elements for frames, faces, joint orientation and bone rotation with respect to the camera's coordinates.

Participants faced each other, and markers were placed on the floor to constrain the participants to a limited area. In addition to the Kinect videos, the recordings included two separate video streams, recorded by conventional video cameras. The Kinect video streams and tracking data were temporally synchronised with audio signals with frames of equal 33ms size. Figure 4.10 depicts the technical set up for the training sessions.

Prosodic properties related to voice quality, fluency, stress and intonation were computed using PRAAT [Boersma and Weenink, 2009]. Kinect body and face tracking data were stored in an `XML` format with elements for frames, faces, joint orientation and bone rotation. Additionally, the resulting media were converted to view, browse and annotate using the Anvil tool[14], see Figure 4.11.

4.6 Annotation and modelling

The ISO 24617-2 dialogue act taxonomy is designed to capture the meaning of dialogue contributions in multiple dimensions, resulting in multi-layered annotations. Nine dimensions are distinguished, addressing information about a certain *Task*; the processing of utterances by the speaker (*Auto-feedback*) or by the addressee (*Allo-feedback*); the manage-

[14] `http://www.anvil-software.org/`

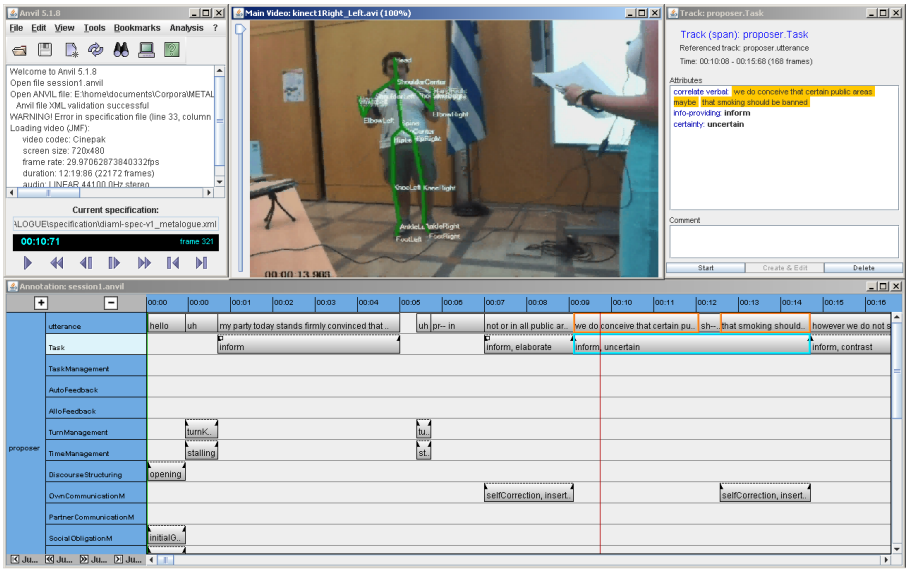

Figure 4.11: Viewing, browsing and annotations of multimodal trainee's behaviour using Anvil.

ment of difficulties in the speaker's contributions (*Own-Communication Management*) or that of the addressee (*Partner Communication Management*); the speaker's need for time to continue the dialogue (*Time Management*); the allocation of the speaker role (*Turn Management*); the structuring of the dialogue (*Dialogue Structuring*); and the management of social obligations (*Social Obligations Management*).

The ISO 24617-2 annotation schema however cannot be expected to be ideal for every kind of dialogue analysis, for every task domain, for every kind of dialogue, and for every annotation purpose. Nevertheless, general principles underlying the design of the schema and the DiAML annotation language are useful for accommodating extensions, modifications, or restrictions of the schema and the annotation language, as the need arises for particular applications. We followed the main design principles and guidelines for schema extension and restriction formulated in ISO 24617-2 standard in Section 12.

For our purposes, we considered one additional dimension (10 in total), *Contact Management*, which is non-core optional in ISO24617-2, since, for example, in debates tutoring sessions managing the contact is an important aspect in such types of dialogues.

Moreover, we introduced 4 additional dimension-specific functions and 1 general-purpose function that are not included in ISO 26417-2, however, defined in DIT^{++} [Bunt, 1999]:

- *Dialogue Act Announcement*, where the speaker makes explicit what kind of dialogue act he/she is going to perform next;
- *Topic introduction*, where the speaker wants to introduce the topic mentioned in the semantic content;
- *Topic shift announcement*, where the speaker wants to change the topic.

- *Preclosing*, where the speaker indicates that he/she plans to end the current dialogue shortly;

- *Threat*, where the speaker states his commitment to perform the action in the manner or with the frequency, described in the semantic content; speaker believes the action to be harmful for the addressee

Additionally, to enable better debate modelling and consistent participant's information state update in system context model we will consider 5 different Auto- and Allo-Feedback levels as defined in DIT^{++}.

Since we aim at developing a Cogntive Tutoring System (CTS), dialogue modelling is concerned with educational dialogues containing tutoring interventions. In order to model tutoring sessions adequately, in Task dimension-specific dialogue acts are considered, like Open Training Session, Suspend Session, Resume Session, etc.

Task Management dimension has been introduced in order to tag dialogue acts dealing with managing the underlying debate and negotiation. The DAMSL annotation scheme also defines this dimension. Dialogue acts in this dimension explicitly address the debate or negotiation process and procedure. This includes utterances that involve coordinating the activities of the two speakers (e.g., "Are you keeping track of the time?", "Let's work on the first issue", etc.), asking for help on the procedures (e.g., "Do I need to state the problem?") or asking about the status of the process (e.g., "Are we done?"). It is important to distinguish between utterances that concern the task management when addressing the task procedures, and discourse structuring when addressing the interactive/dialogue procedures.

Since argumentation structure plays an important role in debates, a set of rhetorical relations, which is left unspecified in ISO 24617-2, is extended. Currently the following set (mainly based on PDTB [Prasad et al., 2008]) is considered but is not exhaustive and will be modified at a later stage of the annotation process, see below.

4.6.1 Annotation design: debates

Planning and preparation of arguments in a debate involves **A**rgument as a general claim , **R**eason(-s) and **E**vidence. This structure is often called **ARE**, see Figure4.6[15].

Good debaters are distinguished by concise clear arguments and try to make their arguments understandable for their addressees. For this purpose, debaters often use linguistic cues such as discourse markers and meta-discoursive acts[16] For example, 'I will talk in favour of ... Because ... Since international research shows...'. Thus, *discourse relations* between two or more *dialogue acts* (argument premises or conclusions) are often marked explicitly by means of discourse markers to support Justification, Motivation, Cause/Result,

[15] See `http://www.slideshare.net/Cherye/advanced-debating-techniques` and [Petukhova et al., 2016a]

[16] [Crismore et al., 1993] define metadiscourse as "linguistic material in texts, written or spoken, which does not add anything to the propositional content but that is intended to help the listener or reader organise, interpret and evaluate the information given", e.g. Shifting Topic, Marking Asides, etc.

Discourse relation	Relative frequency (in %)	Cohen's kappa scores
Elaboration**	28.1	0.67
Evidence**	21.4	0.72
Justify***	16.1	0.76
Condition***	0.7	0.34
Motivation**	1.4	0.48
Background**	0.3	0.18
Cause***	3.4	0.37
Result***	2.2	0.26
Reason*	10.6	0.73
Conclude**	5.7	0.71
Restatement***	10.1	0.76

Table 4.2: Distribution of Inform acts connected by a discourse relation in the corpus (* defined in DPTB; ** defined by Hovy and Maier, 1995; *** in both taxonomies).

Background/Evaluation, Evidence and Circumstance links. Figure 4.6 depicts the most frequently observed argumentation patterns, about 80% of the data follows these patterns. The main claim, i.e. a Statement, is supported by either a Reason or Evidence, and is wrapped up by a Re-Statement in the form of a Summary or Conclusion. For example:

(8) $D1_{21}$[17]: Past anti-smoking campaigns were useless [*Inform*]
$D1_{22}$: I haven't actually seen any of those implemented [*Inform Motivate* $D1_{21}$]
$D1_{23}$: I have personally walked into a store and seen a fourteen years old buying a pack of cigarettes [*Inform Evidence* $D1_{21}$]
$D1_{24}$: Many cases of civil disobedience make this campaign look nice only on paper [*Inform Re-Statement* $D1_{21}$]

In the data, more than 41.4% of the dialogue acts performed by the debaters are Inform acts, which are often connected by discourse relations to form an argument. Small portions of *Set Questions* (3.4%) and *Agreements* or *Disagreements* (1.7%) are observed. Other dialogue acts are concerned with *Turn Management* (22.7%); *Time Management* (21.1%); *Own Communication Management* (7.3%); *Social Obligation Management* (1.2%); and *Discourse Structuring* (10%).

Discourse Relations

Discourse relations were annotated using the annotation scheme designed for the Penn Discourse TreeBank (DPTB) corpus [Prasad et al., 2008]), extended with discourse segment relations from the taxonomy proposed in [Hovy and Maier, 1995]. Table 4.2 presents the types and frequencies of the relations along with the inter-annotator agreement reached annotating each relation type. For relations like Elaboration, Evidence, Justification, Reason, Conclude and Restatement, which are important for the debate argument identification and processing, a substantial agreement has been achieved. The annotated discourse relations

[17] Here and henceforth D*k* stands for Debater *k*; the subscript is the index of the identified dialogue act.

were mapped to those defined in the ISO 24617-8 standard, which was published after all DTC sessions were annotated.

Argumentative Discourse Units

We segmented debates into Argumentative Discourse Units (ADUs), defined as a unit which consists of one or more premises and one conclusion, possibly restated or paraphrased several times by the same speaker. To identify ADUs, we followed the approach proposed by [Peldszus and Stede, 2013], who suggest to first segment into Elementary Discourse Units (EDUs)[18] as minimal discourse building blocks, then establish relationships between two or more EDUs, and combine those into ADUs.

Identifying ADUs, we observed a very frequent pattern[19]: an ADU will mostly start with a simple Inform act and end when an Inform Conclude or Restatement is identified, or before another Inform act is performed by the same speaker which is not involved in any discourse relation, see Figure 4.6 and example (8), or another speaker claims the turn.

Finally, to capture support and attack links between arguments produced by different speakers, we identified explicit and implicit agreement and disagreement dialogue acts signalling support or attack of arguments through the *functional dependence relations* defined in [ISO, 2012] between the detected argument conclusions. For example:

(9) $D1_{47}$;$D1_{1.2}$: The government should launch an effective anti-smoking campaign before it's too late [*Inform*]
$D2_5$;$D2_{2.1}$: The decision to smoke or not is a personal issue and the state shouldn't interfere [*Inform*& *Disagreement* $D1_{47}$] - Attack $D1_{1.2}$
$D7_2$;$D7_{7.1}$: I think public health is one of the most important tasks that the government should perform [*Inform*& *Agreement* $D1_{47}$& *Disagreement* $D2_5$]- Support $D1_{1.2}$/Attack $D2_{2.1}$s

Debater 1 states that an anti-smoking campaign is needed and it is the government responsibility. Debater 2 thinks that smoking is a personal responsibility and the government should not interfere. Debater 7 supports argument 1.2 and thereby attacks the arguments 2.1. These links are modelled as part of the debaters' information states, see [Petukhova et al., 2016a].

4.6.2 Annotation design: negotiations

At the negotiation task level, human-computer negotiation dialogue is often modelled as a sequence of offers. The offers represent participants' commitments to a certain negotiation outcome. In human negotiation, however, offers as binding commitments are rare and a larger variety of negotiation actions is observed, see [Raiffa et al., 2002]. Participant actions are focused mainly on obtaining and providing preference information. A negotiator often states his preferences without expressing (strong) commitments to accept an offer that includes a positively evaluated option, or to reject an offer that includes a negatively evaluated option. To capture these variations, we distinguished five levels of commitment using

[18] EDUs span two dialogue acts connected by a discourse relation.

[19] The inter-annotator agreement between three experienced annotators on this task was very high, 0.87 in terms of Cohen's kappa.

Dialogue Act		Relative frequency (in %)		Dialogue Act		Relative frequency (in %)	
Communicative function	Modality/ Qualifier			Communicative function	Modality/ Qualifier		
propositionalQuestion			2.0	suggest			10.0
checkQuestion			2.2	addressSuggest			1.4
setQuestion			10.3	acceptSuggest			2.0
choiceQuestion			0.6	declineSuggest			1.7
inform $->$			30.3	offer $->$			16.7
...	non-modalised	41.3		...	conditional	28.3	
...	prefer	30.4		...	tentative	35.0	
...	disprefer	3.1		...	final	36.7	
...	acquiesce	3.0		addressOffer			0.6
...	need	2.0		acceptOffer $->$			5.8
...	able	19.0		...	tentative	47.6	
...	unable	1.2		...	final	52.4	
agreement			10.3	declineOffer	tentative		2.0
disagreement			4.1				

Table 4.3: Distribution of task-related dialogue acts in the analysed multi-issue bargaining dialogues.

the ISO 24617-2 dialogue act taxonomy[20] and its superset DIT^{++}[21]: (1) zero commitment for offer elicitations and preference information requests, e.g. by questions; (2) the lowest non-zero level of commitment for informing about preferences, abilities and necessities, e.g. in the form of modalised answers and informs; (3) an interest and consideration to offer a certain value, i.e. suggestions; (4) weak (tentative) or conditional commitment to offer a certain value; and (5) strong (final) commitment to offer a certain value, see Petukhova et al., 2017.

To model negotiation behaviour with respect to preferences, abilities, necessity and acquiescence, and to compute negotiation strategies as accurately as possible, we define several *modal relations* between the modality 'holder' (typically the speaker of the utterance) and the target which consists of the negotiation move (and its arguments), see Lapina and Petukhova (2017). Additionally, to facilitate structuring the interaction and enable participants to interpret partner intentions, dynamically changing goals and strategies efficiently, we defined a set of *qualifiers* attached to offer acceptances or rejections and agreements, tentative or final.

Semantically, dialogue acts correspond to update operations on the information states of the dialogue participants. They have two main components: (1) the *communicative function*, that specifies how to update an information state, e.g. Inform, Question, and Request, and (2) the *semantic content*, i.e. the objects, events, situations, relations, properties, etc. involved in the update, see [Bunt, 2000], [Bunt, 2014a]. Negotiations are commonly analysed in terms of certain actions, such as offers, counter-offers, and concessions, see [Watkins, 2003], [Hindriks et al., 2007]. We considered two possible ways of using such actions, also referred to as 'negotiation moves', to compute the update semantics in negotiation dialogues. One is to treat negotiation moves as task-specific dialogue acts. Due to its domain-independent character, the ISO 24617-2 standard does not define any commu-

[20] For more information see [Bunt, 2009]; visit also `http://dit.uvt.nl/\#iso_24617-2`

[21] `http://dit.uvt.nl/`

Negotiation Move	Relative frequency (in %)
Offer	75.0
CounterOffer	12.4
Exchange	6.6
Concession	1.2
BargainIn	0.4
BargainDown	1.2
Deal	2.4
Terminate	0.8

Table 4.4: Negotiation moves and their relative frequencies in the annotated multi-issue bargaining corpus.

nicative functions that are specific for a particular kind of task or domain, but the standard invites the addition of such functions, and includes guidelines for how to do so. For example, a negotiation-specific kind of $Offer_N$ function could be introduced for the expression of commitments concerning a negotiation value.[22] Another possibility is to use negotiation moves as the semantic content of general-purpose dialogue acts. For example, a negotiator's statements concerning his preference to a certain option can be represented as $Inform(A,B,\Diamond offer(X;Y))$. We specified 8 basic negotiation moves, whose distribution in the analysed data is shown in Table 4.4.

To sum up, the designed negotiation dialogue model accounts for several types of action performed by negotiators: (1) task-related dialogue acts expressing negotiation preferences and commitments; (2) qualified ('modalised') actions expressing participants' negotiation strategies, see Table 4.3; (3) negotiation moves specifying events and their arguments, see Table 4.4; and (4) communicative actions to control the interaction, see Table 4.5. A detailed specification of negotiation update semantics is defined in Petukhova et al. (2017).

Semantic annotations were performed by three trained annotators who reached a good inter-annotator agreement in terms of Cohen's kappa of 0.71 on average, when performing segmentation and annotation simultaneously. In total, the corpus data contains more than 18.000 annotated entities. Annotations are delivered in ISO DiAML format (ISO 24617-2, 2012), `.diaml` files consisting of primary data in TEI-compliant representation, with 24617-2 dialogue act annotations. The collected data and annotations is part of the Metalogue Multi-Issue Bargaining (MIB) corpus (Petukhova et al., 2016) which is released through LDC.[23].

4.6.3 Querying additional dialogue resources

For accessing existing annotated corpora, first, a mapping to the annotation language defined by the standard should be achieved. In the past, there have been several attempts to perform a comparable task by comparing and mapping existing dialogue act annotation schemes. The most recent one was performed within the work on the ISO standard 24617-2. The ISO dialogue act annotation scheme was mapped to 18 existing dialogue act annotation schemes,

[22] Negotiation 'Offers' may have a more domain-specific name, e.g. *Bid* for selling-buying bargaining.

[23] Please visit `https://catalog.ldc.upenn.edu/LDC2017S11`

ISO 24617-2 dimension	Relative frequency (in %)
Task	47.6
Task Management	10.3
AutoFeedback	18.7
AlloFeedback	2.3
Turn Management	6.6
Time Management	6.6
Discourse Structuring	4.6
Own Communication Management	2.1
Partner Communication Management	na
Social Obligation Management	1.2

Table 4.5: Distribution of dialogue acts per ISO 24617-2 dimension in the multi-issue bargaining corpus.

see Petukhova (2011) and the informative Annex F 'A survey and analysis of dimensions and communicative functions in existing annotation schemas' of the standard.

Analysing annotation schemes and data representations, and their compatibility with ISO 24617-2, at least four very important issues need to be taken into consideration: (1) the multifunctionality of dialogue utterances; (2) the way a dialogue is segmented into meaningful units; (3) the relations between segments; and (4) the qualification of communicative functions. Existing annotation schemes take different points of view on these issues. For the purpose of querying we largely ignore differences in segmentation, and use the segmentation used in the corpus.

The ISO-compatibility of an annotation scheme can be considered at many levels (Bunt et al., 2013). One possibility is to take the communicative function tags used in a given annotated corpus and replace them by ISO tags. Since there is no one-to-one correspondence between tags, this is mostly not a simple matter, but in fact requires the re-expression of the information that is captured by the corpus annotations in terms of concepts defined in the ISO standard. Table 4.6 shows how the functional tags for information-giving and information-seeking acts in the DAMSL, SWBD-DAMSL, AMI, HCRC Map Task, and ISO 24617-2 annotation schemes are related.

We systematically compared the MapTask, AMI and ISO 24617-2 annotation schemes by inspecting the definitions as well as examples in annotation guidelines and annotated corpus data. Additionally, four AMI dialogues (3,897 utterances) and eight MapTask dialogues (1,728 utterances) were re-annotated according to ISO 24617-2.

A big collection of dialogues constitutes the **HCRC MapTask**[24] corpus, consisting of 128 dialogues where one participant plays the role of an instruction-giver while the other participant, the instruction-follower, navigates through the map. The dialogues are transcribed and annotated for a wide range of behaviours, e.g. prosodic and syntactic units, gaze direction, conversational moves, etc. The HCRC MapTask annotated corpus is available in NXT format. Moreover, MapTask's underlying idea was so successful that dialogues for a comparable task (map-searching) has been collected in many languages other than English:

[24] http://www.hcrc.ed.ac.uk/maptask/

ISO(*Qualifier* or Relation)	**DAMSL**	**SWBD-DAMSL**	**AMI**.Relational tag	**HCRC MapTask**
Inform Inform(*Uncertain*) Inform(*Certain*) Inform(Explanation) Inform(Clarification)	(Re)Assert Other Statement (Re)Assert	Statement: Statement-opinion Statement-non-opinion	Inform Inform.Uncertain Inform	Statement Explain Clarify
Agreement	Agreement:Accept Accept-part	Agree Accept Accept-part	Inform.Positive -	Reply-y -
Agreement(*Uncertain*)	Maybe	Maybe	Inform.Uncertain	-
Disagreement	Reject	Reject Disprefered responses	Inform.Negative	Reply-n
Disagreement(*Uncertain*)	Maybe	Maybe	Inform.Uncertain	-
	Reject-part	Partial Reject	-	-
Correction		-	Inform.Negative	-
Set answer	Answer	Answer	Inform	Reply-w
Prop. answer	Answer Answer	Yes-answer No-answer	Inform	Reply-y; Reply-n
Prop. answer(*Negative*)	Answer	Negative non-no answer	Inform.Negative	
Confirm	Answer	Yes-answer	Inform.Positive	Reply-y
Disconfirm	Answer	No-answer Disprefered responses	Inform.Negative	Reply-n
Set question	Info-Request	WH-question	Elicit Inform	Query-w
Propos. Question	Info-Request	YN-question	Elicit Inform	Query-yn
Check-Question	Infor-Request+Assert	Declarative YN-question Declarative WH-question Tag-Question	Elicit Inform	Check
Choice question	Info-Request	OR-question Or-clause	Elicit Inform	Query-w
Question	Info-Request	Open Questions	Elicit Inform	
Question	Forward-Looking	Rhetorical questions	Elicit Inform	

Table 4.6: Information transfer (providing and seeking) communicative functions in the ISO24617-2, DAMSL, SWBD-DAMSL, AMI and HCRC MapTask annotation schemes. From Petukhova et al. (2014).

German MapTask (Hamburg MapTask corpus[25], French MapTask corpus[26] (MAPTASK-AIX), Italian MapTask (Grice and Savino, 2003[Grice and Savino, 2003]), and many others.

The **AMI** corpus[27], collected in a large-scale EU project, contains 100 hours of transcribed and annotated meeting conversations (in English) where the participants (usually four) play different roles in a fictitious design team. The annotated corpus is also available in NXT format.

The above mentioned dialogue corpora differ in (1) underlying task (instructing a map search, decision making, selling-buying negotiations and a collaborative design task); (2) number of dialogue participants (two- or multy-party); (3) communication channels and modalities (computer-mediated typed, face-to-face spoken interactions, and spoken interaction without visual contact); and (4) annotated phenomena and annotation scheme used.

Multifunctionality and multidimensionality

As discussed in Section 2.1, the ISO 24617-2 annotation scheme is highly multidimensional, supporting multifunctional analysis by allowing the assignment of multiple dialogue act tags to a dialogue segment. The ISO 24617-2 taxonomy of communicative functions distinguishes 9 dimensions, taken from the DIT^{++} taxonomy (Bunt, 2009).

Other schemes propose tagsets which as a rule are fairly simple, and are mostly used to code dialogue utterances with a single tag. HCRC MapTask defines such a one-dimensional scheme with 12 mutually exclusive dialogue act tags, see Anderson et al., 1991.

Again other schemes, while allowing a single dialogue act label to be assigned to each dialogue segment, have additional tags that can be added to the main label in order to describe the meaning more accurately. AMI is one such scheme which has additional layers and relational tags. For instance, an additional layer of so-called 'reflexive' acts allows labelling the type of semantic content by specifying whether a dialogue contribution is about the meeting task or about managing the task. Further, the AMI annotation scheme has relational tags to indicate relations between dialogue units. For example, INFORM which can be combined with 4 relation tags: POSitive, NEGative, PARTial and UNCertain. This allows to annotate several types of answers, e.g. positive or negative answer, or positive uncertain answer, etc. It does not allow, however, to differentiate between, for example, a confirm, an agreement and a positive propositional answer, or between those of accept request, accept suggestion and accept offer, which are not concerned with the exchange of information in propositional form, but address the performance of actions.

[25] http://www1.uni-hamburg.de/exmaralda/files/z2-hamatac/public/index.html

[26] http://crdo.up.univ-aix.fr/voir_depot.php?lang=en&id=732&prefix=sldr

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

Relations between dialogue units

ISO 24617-2 distinguishes three types of relations between dialogue units: functional relations (like question-answer), feedback relations and rhetorical relations, as described in Section 2.1. In AMI, annotators should consider whether a unit expresses a response to something, and so, indicate that by adding a link. For instance, an answer is linked to a question (as 'source' of the link; the answer as 'target').

As for rhetorical relations, AMI has a separate scheme to capture argumentation structure, however, not relating dialogue acts but segments.

HCRC MapTask does not explicitly mark any relations between dialogue units. The functional labels *Explain* and *Clarify* are defined to say that the current speech act explains or clarifies something.

Communicative function qualifiers

ISO 24617-2 defines a set of qualifiers to enable more precise description of the speaker's intention with respect to certainty, conditionality and sentiment. Some dialogue act taxonomies pay attention to these phenomena. For instance, DAMSL and DAMSL-based schemes distinguish such functions as Maybe, Reject-Part or Accept-Part. AMI uses the relational tags POSitive, NEGative, PARTial, or UNCertain to classify the type of a relationship. Emotions are also annotated in AMI, however these labels are assigned directly to the (verbal or nonverbal) behaviour of a participant and are not tied with dialogue act annotation. HCRC MapTask does not capture these phenomena.

Tag correspondences

In this section we present the mappings, based on both theoretical and empirical considerations, between the AMI, HCRC Map Task, and ISO 24617-2 annotation schemes.

A first observation is that there are very few one-to-one correspondences between function tags. 'Instruct' in HCRC MapTask and ISO 24617-2 is an example. There are even fewer many-to-one functional tag correspondences from AMI or MapTask to ISO 24617-2, also if we chose a more general ISO tag. For example, AMI's Elicit-Inform, Elicit-Assessment, Elicit-Comment-Understanding and Elicit-Offer-or-Suggestion may be mapped to ISO's general Question tag. Upon analysis and re-annotation it turns out that of the dialogue acts with these functions, Elicit-Inform and Elicit-Offer-or-Suggestion mostly address the Task dimension, while Elicit-Assessment and Elicit-Comment-Understanding are mostly concerned with feedback elicitation. The remaining Elicit-Offer-or-Suggestion maps in about 50% of the cases to the ISO tag 'Question' and in 50% to 'Request'.

In view of the highly multidimensional and detailed nature of the ISO annotation scheme, the most common mapping to that scheme is one-to-many. This is the source of most of the problems for automatically mapping between the annotations in the AMI and MapTask corpora and ISO equivalents. For example, Inform in the AMI corpus may correspond to Inform, Answer, Agreement, Disagreement, and several kind of Accept and Reject tags

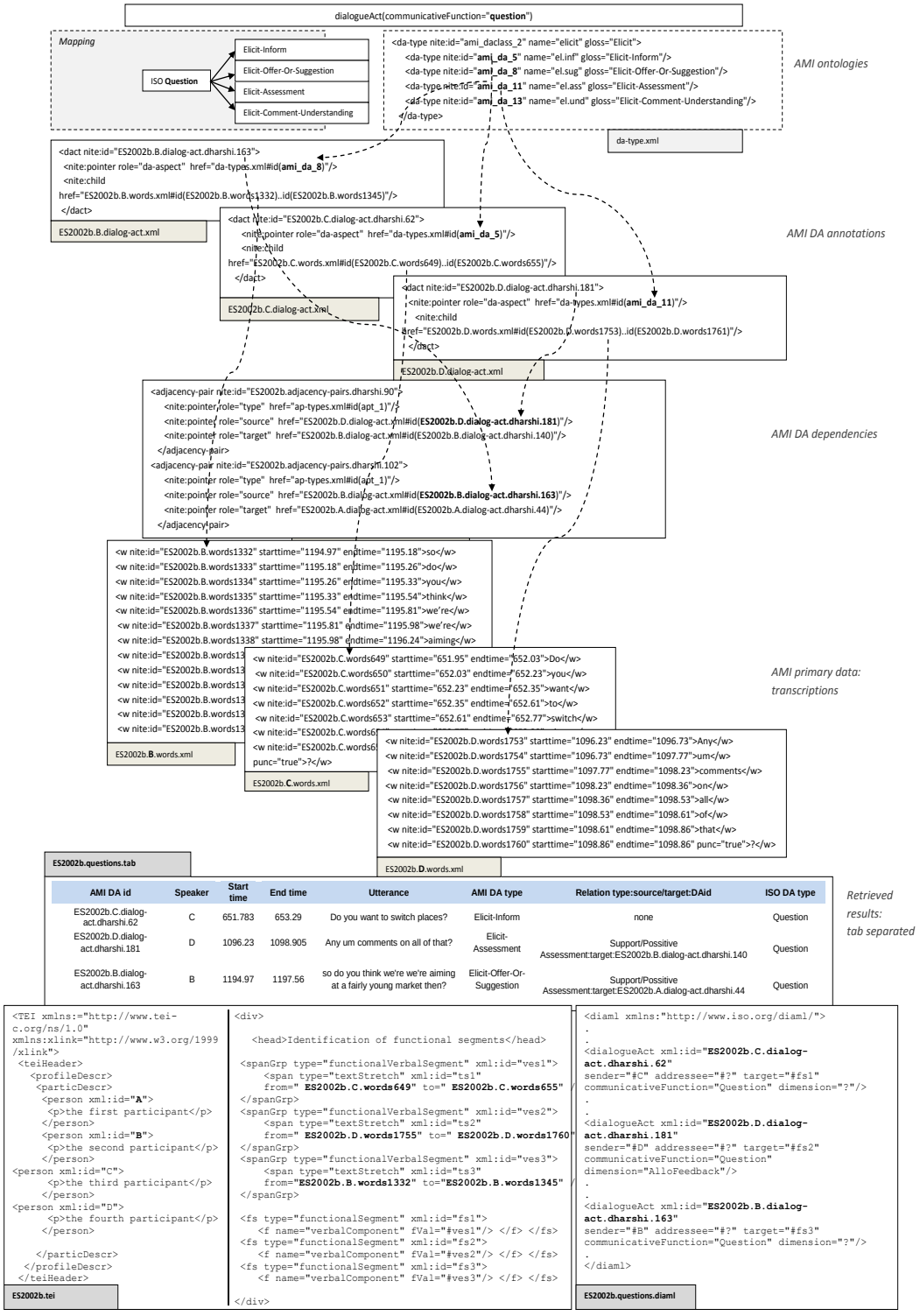

Figure 4.12: Example of Questions retrieval from the AMI meeting corpus and representing them in DiAML.

AMI	Previous AMI tag	AMI relational tag	ISO functional tag
Inform	Elicit Inform	POSitive or NEGative	Answer
	Inform	POSitive	Agreement
		NEGative	Disagreement
	Elicit-offer-or-suggestion	POSitive	Accept Request
			Answer
		NEGative	Decline Request
			Answer
		-	Address Request
			Answer
	Elicit-Comment-Understanding	POSitive	Positive AutoFeedback
		NEGative	Negative AutoFeedback
	Elicit-Assessment	POSitive	Positive AutoFeedback
		NEGative	Negative AutoFeedback

Table 4.7: One-to-many mapping between AMI Inform and corresponding ISO functional tags. From Petukhova et al. (2014).

defined in the ISO standard. To be able to differentiate between these we take the functional tag of the preceding segment and the AMI relational tag into account. If the previous tag was Elicit Inform, then the AMI Inform is mapped to Answer, if the AMI preceding tag was Inform, we map to Agreement if a POSitive relation tag was assigned, and to Disagreement if a NEGative tag was assigned (see Table 4.7).

Querying corpora through DiAML

AMI, HCRC MapTask and ISO 24617-2 annotations are in stand-off form, i.e. the representations are stored in separate files, linked to the primary data, typically using separate files per dialogue and per speaker. If a specific dialogue act (or type of dialogue act) is to be retrieved, we go through multiple annotation files and collect the relevant information. For example, to extract all instances of Question acts in AMI corpus data, the query will be searching for matches for the tags Elicit Inform, Elicit-Assessment, Elicit-Comment-Understanding and Elicit-Offer-or-Suggestion. Figure 4.12 illustrates the query processing and the retrieval workflow. The dialogue act types `da-type` are specified in AMI ontology files (da-type.xml for dialogue acts) where the unique identifier that is the value of `@nite:id` is assigned to each of them, e.g. `ami_da_5` for Elicit-Inform, `ami_da_8` for Elicit-Offer-or-Suggestion, etc. Having this information, we search the dialogue act annotation files, e.g. ES2002b.B.dialogue-act.xml, where ES2002 is a meeting id, b means that it was the second dialogue with these participants, B stands for the speaker who plays the project leader role). Each dialogue act has a unique `@nite:id` identifier as well, which helps to find all other information in AMI data that points to this dialogue act, e.g. adjacency pairs. We collect the primary data that the annotation is attached to. Each identified `<dact>` element is linked to words produced by the corresponding speaker. The start and end words are indicated, for example, as `href="ES2002b.D.words.xml#id(ES2002b.D.words1753)..id(ES2002b.D.words1761)"`. Since we know that AMI does not allow dialogue segments to be discontinuous, we compile the wording of the corresponding utterance by taking every word between start and end token including the former and the later ones, e.g. 'Any um comments on all of that?' for the example in Fig. 4.12. Each

word element <w> in the transcription files has @starttime and @endtime as attributes. The start time of the first token and the end time of the last token of the corresponding dialogue act are then used to compute the utterance time stamps. The figure also shows how the metadata and the primary data can be represented in TEI format and the dialogue act annotation in DiAML.

Some annotations in AMI bypass dialogue act annotations, e.g. rhetorical relation for argumentation structure, disfluencies, etc., and are attached directly to the primary data. To retrieve this information, the workflow would be slightly different. For instance, we need to start with those annotation files, e.g. ES2002a.A.argumentstructs.xml with a word span as @nite:child. Subsequently, we check whether the same word segments are marked for dialogue acts, and compute rhetorical relations between the identified dialogue acts.

Other AMI annotations that are not part of dialogue act annotations, but are relevant to determine, are those for emotions. Emotion tags are assigned in AMI to multimodal data, e.g. words, focus of attention signals, head movements and gestures. In order to relate this information to dialogue acts, time stamps need to be taken into account to identify a multimodal utterance.

There are at least two possible ways of querying existing annotated corpora using DiAML. One way is to transform corpora which are in XML format into DiAML compliant format, and subsequently query these data using XQuery or XPath, designed to query XML data. For example, the XPath query to retrieve all Questions from the AMI data would be:

(10)
```
/AMI-data/*.diaml/
dialogueAct/
[communicativeFunction="question"]
```

or XQuery

(11)
```
for $x in doc("*.diaml)/AMI-data
where
$x/communicativeFunction="question"
order by $x/starttime
return $x/dialogueAct
```

To define a query in the ways shown above, knowledge of the corpus specific annotation scheme and its translation into ISO 24671-2, as well as of the DiAML structure is required.

The second approach is to define a DiAML query format and, provided a good (not necessarily perfect) mapping to ISO 24617-2 exists, use this to directly retrieve desired information from annotated data. Both approaches are valid. The first one presents a standard way of querying XML data. The second approach is a more straightforward and flexible way of DiAML oriented querying of dialogue act annotated data. Since it closely relates to DiAML specification, there is no need to know details of different annotation schemes and their annotation formats. For the DiAML-oriented querying we designed an interface presented in Figure 4.13.

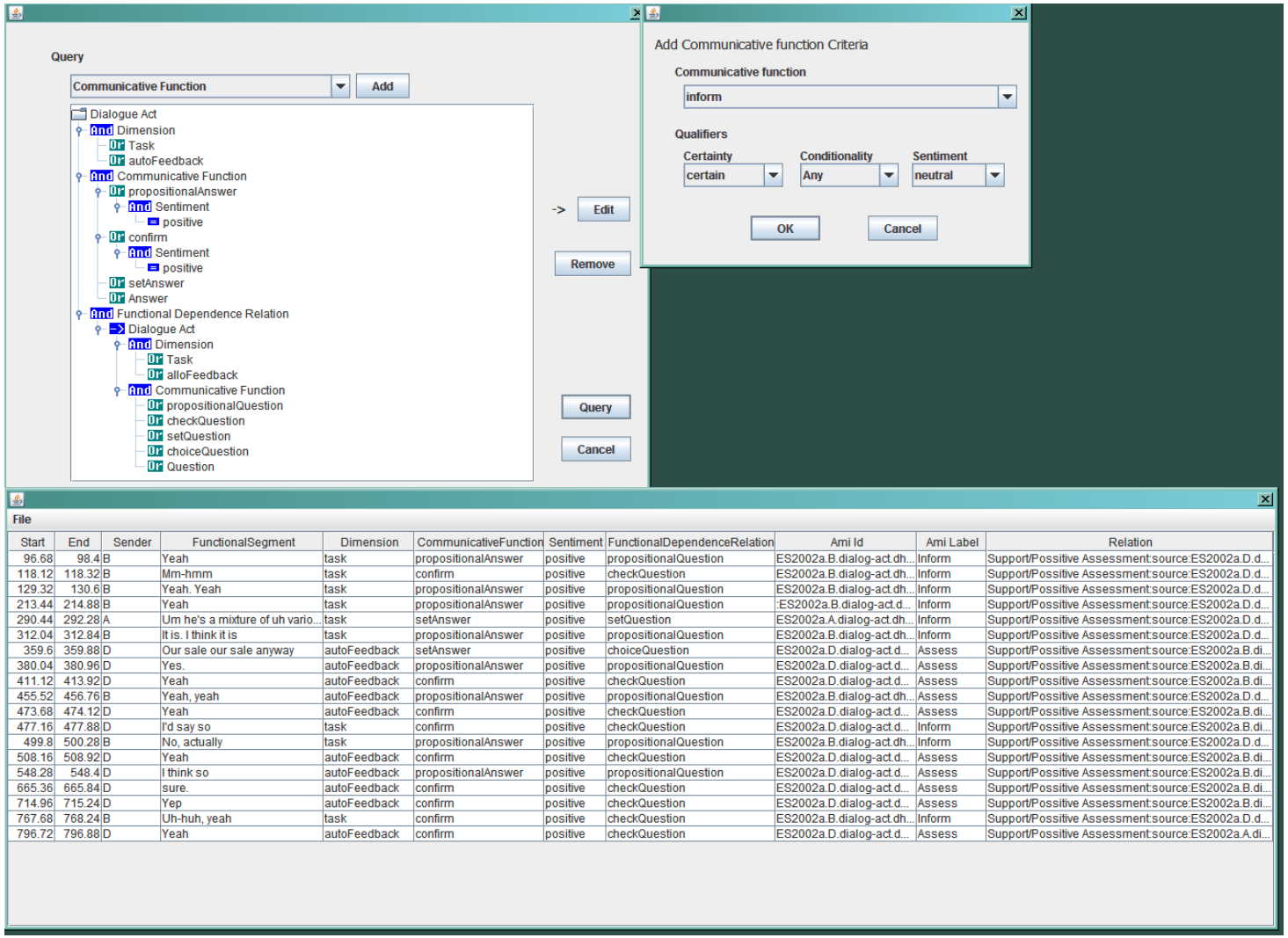

Figure 4.13: Interface for DiAML-based querying of annotated dialogue corpus (example on AMI data).

DiAML query for	Percentage of instances retrieved per query	
	HCRC MapTask	AMI
SetQuestion	2.9	2.3
PropositionalQuestion	7.1	5.8
CheckQuestions	7.1	3.3
SetAnswer	2.4	3.9
PropositionalAnswer	4.3	9.8
Inform	7.8	11.7
Instruct	26.8	0.3
Suggest	0.0	10.1
PositiveAutoFeedback	15.7	20.5
FeedbackElicitation	4.7	0.7

Table 4.8: Retrieval performance on HCRC MapTask data.

Results and their validation

Table 4.8 presents the retrieval results when querying HCRC MapTask and AMI corpora (per ISO functional tag) in terms of relative frequency in the given corpus data.

The results have been evaluated in terms of precision and recall. While precision is the fraction of retrieved instances (i.e. utterances with the queried tag) that are relevant to the query and indicates the correctness of the retrieved results, recall is the fraction of the instances that are relevant to the query that are successfully retrieved and indicates the completeness of the retrieved results (i.e. the lower recall the more relevant instances are missed by the query). For this, we compared the retrieved results for each dialogue act type with manually produced reference annotations. Table 4.9 presents the evaluation results for HCRC MapTask and AMI data for each ISO dialogue act type occurring in the corpus data.

The results show reasonably high precision and recall for all types of dialogue acts, except for Feedback Elicitation utterances in case of the HCRC MapTask corpus which cannot be easy mapped. Such utterances correspond to most of the HCRC MapTask 'aligns' and 'ready' utterances, and sometimes to 'query-w' and 'query-yn' utterances, where no clear-cut distinction can be made without taking more complex dialogue properties into account (e.g. larger dialogue history in combination with the wording of dialogue contributions from the left context, where the latter was ignored in the experiments reported here).

4.7 Implementation and testing

The Virtual Negotiation and Debate Coaches "hear" and "see" a wide range of signals, interpret them and act as an observer, as a negotiation partner, and/or as a tutor.

The speech signals and tracking data serve as input for further processing. The Kaldi-based ASR [Povey, 2011] was trained on 759 hours of data[28], achieving a performance of 34.4% Word Error Rate (WER), see [Singh et al., 2017].

For semantic interpretation, the ASR output was used for the event, arguments and modality classification, and communicative function recognition. Conditional Random

[28]The following resources were used: the Wall Street Journal WSJ0 corpus, HUB4 News Broadcast data, the VoxForge, the LibriSpeech and AMI corpora.

Query for	HCRC MapTask		AMI	
	Precision	Recall	Precision	Recall
SetQuestions	87.8	92.0	88.5	91.3
PropositionalQuestions	81.1	65.8	75.8	68.5
CheckQuestions	75.2	67.6	73.2	56.4
SetAnswer	65.0	59.5	77.5	54.2
PropositionalAnswer	73.2	69.1	77.8	66.8
Confirm	71.4	62.5	71.7	47.3
Inform	83.5	64.4	80.5	79.8
Inform Elaborate	na	na	79.4	72.1
Inform Explain	81.3	72.6	66.7	63.1
Inform Clarify	74.8	29.6	73.7	47.8
Request/Instruct	80.8	92.1	75.8	93.7
Suggest	na	na	65.6	60.5
PositiveAutoFeedback	72.1	68.3	95.1	89.3
FeedbackElicitation	52.2	21.8	78.8	57.1

Table 4.9: Retrieval performance for HCRC MapTask and AMI data per query.

Fields models [Lafferty et al., 2001] were trained to predict negotiation moves which specify events and their arguments, as well as their boundaries in the ASR 1st-best string. The classifier predicts three types of classes: negotiation move (event), issue and preference value (event participants, i.e. semantic roles) obtaining an F-score of 0.7 on average. The obtained interpretation is of type $offer(ISSUE = X; VALUE = Y)$. The Support Vector Machine [Vapnik, 2013] modality classifiers show accuracies in the range between 73.3 and 82.6% [Petukhova et al., 2017a]. The obtained interpretation of a modalised negotiation move stating preference is represented as $\Box offer(ISSUE = X; VALUE = Y)$.

The manually ISO 24617-2 annotated Debate Trainee Corpus [Petukhova et al., 2017b] and Multi-issue Bargaining Corpus [Petukhova et al., 2016b] served as initial training data for communicative function classifiers. Additionally, the in-domain data was enriched with those from the MapTask [Anderson et al., 1991], AMI[Carletta, 2006], and Switchboard-DAMSL [Jurafsky et al., 1997] corpora. F-scores ranging between 0.83 and 0.86 were obtained in SVM-based clasification experiments, which corresponds to state-of-the-art performance, see [Amanova et al., 2016].

Kinect tracked data is used to detect hand/arm co-speech gestures[29] and their types, e.g. beats, adaptors, iconics, deictics and emblems. SVM and Gradient Boosting [Friedman, 2002] classifiers were trained and achieved F-scores of 0.72 [Petukhova et al., 2017c]. The motion interpretation component related to hand/arms position detection of the designed Presentation Trainer ([Van Rosmalen et al., 2015, Schneider et al., 2015a]) is integrated into the VDC system.

To obtain context dependent interpretation, dependence relations were computed from the dialogue history stored in the linguistic context of the Dialogue Manager (DM), see next Chapter. The discourse relations recognition is important for discourse-based argument structure recognition [Petukhova et al., 2017b]. The SVM-based classifier yielded F-scores of 0.54 on a coarse 3-class task (Contingency, Evidence, No-Relation) and 0.46 on a fine-grained 7-class task (Justification, Reason, Motivation, Exemplification, Explanation,

[29] Co-speech gestures are visible hand/arm movements produced alongside speech and are interpretable only through their semantic relation to the synchronous speech content.

Exception and No-Relation).

At the semantic fusion level, verbal, prosodic and motion tracking information is combined to obtain complete multimodal dialogue act interpretations, consumed by the Dialogue Manager (DM). The DM, designed as a set of processes (threads), receives data, updates the information state and generates the next action(-s) of the system, see Chapter 5. The DMs in the VNC and in the VDC applications differ, since the two systems have different roles and tasks. As a Debate Coach, the system observes debaters' behaviour, evaluates it on criteria related to (1) how convincing is a debater's argumentation; (2) how well are debate arguments structured; and (3) how well is an argument delivered, and generates real-time 'in-action' feedback, see [Petukhova et al., 2017b]. As a Negotiation Coach, the system performs as a negotiation partner and as a Tutor providing feedback on a trainee's negotiation behaviour. The DM monitors and reasons about the overall state of the negotiation or debate task. The DM takes care of feedback and dialogue control actions concerning contact and social obligations management, as well as recovery and error handling actions.

While task-related dialogue acts are application- and user-specific, in a shared cultural and linguistic context, the choices concerning the frequency of dialogue control actions and the variety of expressions are rather limited, notably for feedback and turn management. Models of dialogue control behaviour once designed can therefore be applied in a wide range of communicative situations. This was one of the main motivations behind the multi-agent DM architecture (Figure 5.1) where task-related and dialogue control agents/managers are separated. When integrated into different dialogue systems mostly parts of Task Managers are replaced, while other parts were largely re-used without significant changes.

Given the dialogue acts provided by the DM, the Fission module generates system responses using pre-defined templates for each dimension, splitting content into different modalities: Avatar[30] and Voice (TTS[31]) actions are generated for the system in partner mode, and visual feedback as tutoring actions. The latter include feedback on presentational aspects and cooperativesness level, visualised by happy and sad face emoticons. At the end of each negotiation and debate session, summative feedback is generated about several aspects of the trainee performance and learning progress. More details on the system implementation and evaluation is provided in Chapter 6.

4.8 Corpus evaluation and deployment

Full session recordings, system recognition and processing results, and the generated dialogue system responses were logged and converted to `.anvil` format for post-processing with the Anvil tool. This tool allows user-defined coding schemes, offering various tier relationships and controlled vocabularies. The tiered format is convenient for transcrip-

[30] Commercial software of Charamel GmbH has been used, see [Reinecke, 2003]

[31] Vocalizer of Nuance, `http://www.nuance.com/for-business/text-to-speech/vocalizer/index.htm`, was integrated.

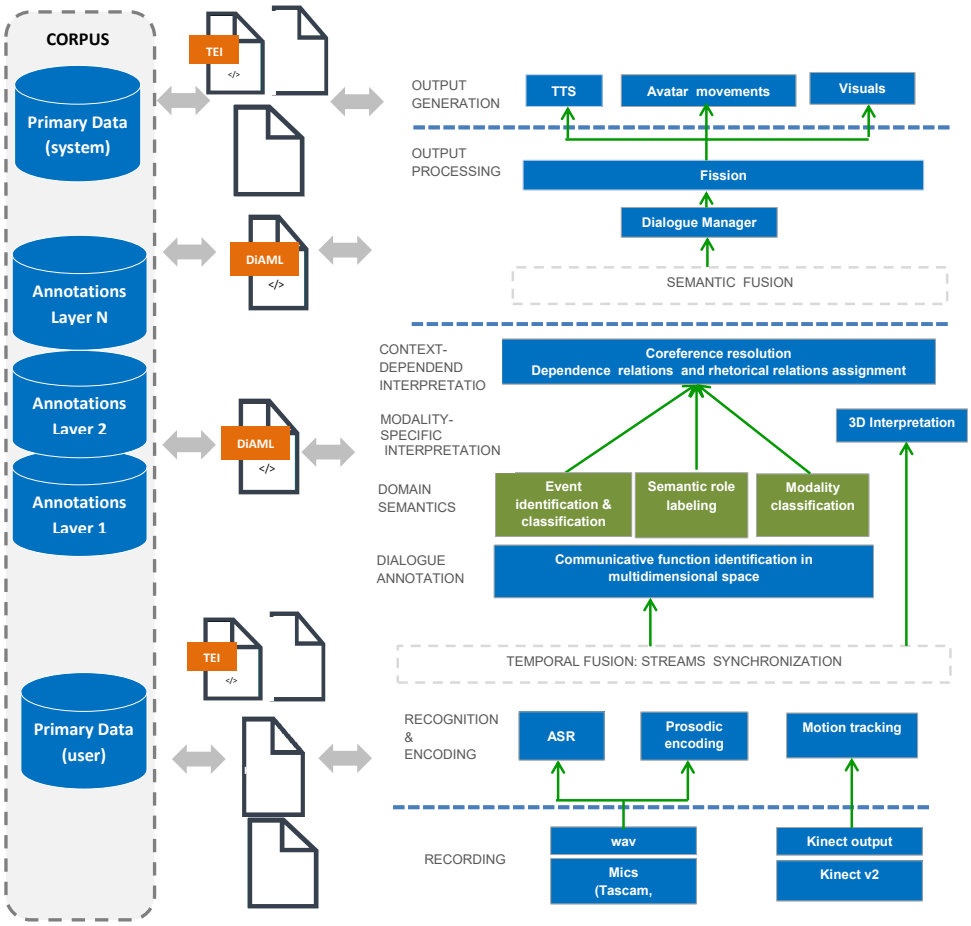

Figure 4.14: Overall corpus creation and system architecture. From bottom to top, signals are received through input devices, and processed by tailored modules. After annotation concerned with communicative function classification, domain-specific semantic and modality-specific interpretations, context-dependent representations are fused and passed to the Dialogue Manager for context model update and system response generation. The system output is rendered in different output modalities. The generated behaviour is written back to the corpus as primary data and representations as annotations, and proposed for editing by human annotators and system module re-training.

tions and annotations in multiple modalities and dimensions. Stretches of communicative multimodal behaviour are marked up with multiple tags, especially when the various tags provide functional information relating to a particular dimension of interaction, such as feedback, turn taking, or time management, see Petukhova and Bunt (2010b), Bunt et al. (2012b) and Petukhova (2011). Annotations are stand-alone and performed using the Anvil specification designed for ISO 24617-2[32].

The Anvil functionality was extended to allow experimenting with variations in system behaviour by tuning, replaying and repairing it. Corrected transcriptions and annotations served: (1) evaluation, measuring inter-annotator agreement to assess corpus data usability, and module-based evaluation contrasting system and human performance on all processing tasks; (2) revision of scenario, requirements and data models; and (3) re-training modules on more and better data in order to improve the system performance.

Two resulted corpora are evaluated and deployed when designing the Virtual Negotiation Coach and Virtual Debate Coach applications. They are documented and either released or are in preparation to be released to the research community - the Multi-Issue Bargaining Corpus[33] and the Debate Trainee Corpus[34]. Figure 4.14 summarizes the overall corpus and system development framework.

4.9 Summary

In this chapter we briefly discussed different approaches to the creation of interoperable dialogue corpora annotated with dialogue act information: (1) creation of new annotated corpora using ISO 24617-2 standard taxonomy; (2) re-annotation of existing annotated data with the ISO standard; and (3) conversion of existing annotated data into ISO 24617-2 compliant format through mapping between dialogue act annotation schemes.

We proposed the continuous corpus creation (D3C) methodology. This method serves multiple purposes. A dialogue corpus is seen as a shared dynamic and evolving repository of information necessary for analysis and modelling of interactive dialogue behaviour, and for implementation, integration and evaluation of the dialogue system. The corpus is created and enriched with every interaction of the user with the developed dialogue system. This not only allows the collection, monitoring, and analysis of real user data by humans and automatically applying advanced data processing algorithms, but also to develop, test, evaluate and re-train system components on the fly which can evolve and improve over time. The key enabler are standard data models allowing all system components to exchange information based on standard data formats, viz. TEI and DiAML.

[32] An example specification is available at `http://www.anvil-software.org/data/diaml-spec-v0.5.xml`

[33] `https://catalog.ldc.upenn.edu/LDC2017S11`

[34] See Petukhova et al., 2018

Type	Content	Format	Comment
Debaters minimal goals cards	4 rounds	pdf	defined for Proponent and Opponent
Negotiators preference cards	9 negotiation cases	html for web-presentation stand-alone GUI (Java)	defined for each negotiator
Metadata	participants (id, native language sex, age at collection)	xml, TEI compliant	generated form participants forms
Signals	sound recordings	mono, 96000Hz sample rate 24-bit sample format	1 channel per speaker
	wav files	mono, 16-bit sample format	cut per speaker/per turn
	Kinect tracking	xml, 30 frames per second	tracked per speaker
		avi videos	recorded per speaker
Automatic Speech Recognition	turn (id, start, end, string)	plain text	automatic
Transcriptions	turn (id, start, end, string)	plain text	manual
	utterance (id, start, end, string)	xml, TEI compliant	automatic
	functional segments (id, start, end, pointers)	xml, TEI compliant	automatic
DA annotations	dialogue act (sender, dimension, communicative function, qualifier functionalDependenceRelation feedbackDependenceRelation) rhetoricalLinks	Anvil and DiAML	manual
Negotiation Moves	events, arguments, links	Anvil and DiAML	manual

Table 4.10: Metalogue Multi-Issue Bargaining and Debate Trainee corpora overview.

The D3C approach has been extended with the development of a query format to access existing annotated corpora through a mapping to the annotation language defined by the standard. We applied the proposed approach to two different dialogue corpora, annotated with different dialogue act annotation schemes. This demonstrates the portability of the approach. The query format based on DiAML provides an attractive solution, since it can be applied to query many different annotated dialogue resources provided a good mapping between tagsets is achieved. There are other annotated dialogue resources that can be added to the collection of interoperable resources. 18 existing dialogue act annotation schemes have been mapped to ISO 24617-2 (see informative Appendix F of the standard). In particular, DAMSL and its existing variants such as Coconut-DAMSL, SWBD-DAMSL and MRDA should be added to this pool. Moreover, dialogue resources for other languages than English can be explored. The designed querying tool has been used to retrieve additional annotated data enriching the training set for automatic cross-domain dialogue act classification using machine-learning algorithms, see Amanova et al. (2016).

During the data collection experiments two new corpora have been created - the Metalogue Multi-Issue Bargaining (MIB) corpus (Petukhova et al., 2016) and the Metalogue Debate Trainee Corpus (DTC, Petukhova et al., 2018). The corpora have speech signals captured with two headset microphones and transcribed both automatically and manually. Additionally, the DTC corpus contains Kinect tracking data of visual body movements. The MIB corpus consists of 24 dialogues of a total duration of about 2.5 hours comprising about 2.000 speaking turns, 3.650 functional segments and about 10.000 tokens. The DTC corpus contains 12 debate sessions with a duration of 2.5 hours, comprising 400 arguments.

Seven types of semantic annotation were performed by three trained annotators reaching a good inter-annotator agreement in terms of Cohen's kappa of 0.71 on average, both on segmentation and annotation tasks. The following entities were identified and annotated:

- Dialogue acts in the 9 ISO dimensions[35].
- Discourse structuring acts according to DIT^{++}
- Contact Management acts according to DIT^{++}
- Task Management dialogue acts
- Negotiation moves as defined in Petukhova et al. (2016)
- Rhetorical relations appying the ISO 24617-8 discourse relation set (Bunt and Prasad, 2016)
- Disfluencies in speech production as defined in Besser (2016)

In total, both corpora contain about 19.000 annotated entities. Table 4.10 provides an overview of the corpora contents. Annotation files of `.diaml` type consist of a TEI-compliant primary data representation and ISO 24617-2 dialogue act annotations attached to the primary data. Metalogue specific elements based on DiAML types are defined inside the `metalogue` namespace in `MlogContent.xsd` scheme file. `MlogCorpus.xsd`

[35] See `http://dit.uvt.nl/\#iso_24617-2`

combines TEI element from `tei` namespace (containing TEI-compliant primary data representation) and `diaml` element from `metalogue` namespace (containing metalogue specific annotation) into the `dialogueSession` element. Using several `xsd` files enables the development of a single `MlogContent.xsd` scheme that declares all data structures common for all of the stages and developments of the work presented here (developments such as collecting data, automatically generating parts of the software code, publishing of corpus data). Thus, guaranteeing (albeit under ordinary conditions) that the structure of data remains consistent overall and throughout all of the mentioned developments. All Metalogue corpora related XSD schemes are provided with the corpus release. As for TEI schemes, we refer to the TEI standard specific documentation at `http://www.tei-c.org/index.xml`. Both resources are released to the community for research purposes under the LDC and ELRA licenses.

Chapter 5

Multi-Agent Dialogue Management

This chapter provides a detailed description of the Multi-Agent Dialogue Manager, a core system module. The manager is designed within the ISU framework, in particular the context of the DIT. We specify the DIT multidimensional context model in more detail and extend the dialogue act update semantics with domain-specific semantic content, i.e. debate and negotiation semantics. As Task Agents two different cognitive models are integrated - Debate Agent and Negotiation Agent. Further, task-related actions handled by cognitive agents are separated from dialogue control actions enabling the application of sophisticated models along with a flexible architecture in which various (including alternative) modelling approaches can be combined. Important high-level error handling strategies are defined. Task Agents are evaluated comparing human-human and human-agent performance on the same task.

Introduction

This Chapter presents the dialogue management approach that incorporates cognitive task models into Information State Update (ISU) based dialogue management as a part of a multimodal dialogue system. Such integration has important advantages. The ISU methodology, presented in Section 2.5.4, has been applied successfully to a large variety of interactive tasks, e.g. information seeking (Keizer et al., 2011), human-robot communication (Peltason and Wrede, 2011), instruction giving (Lauria et al., 2001), and controlling smart home environments (Bos et al, 2003)). Several ISU development environments are available, such as TrindiKit (Larsson and Traum, 2000) and Dipper (Bos et al., 2003). The ISU approach provides a flexible computational model for understanding and generation

The work reported in this chapter takes as a starting point the approach and Dialogue Manager architecture proposed in Malchanau et al. (2015) and extends them (see also Malchanau et al., 2019), for which I performed the research, in close collaboration with my co-authors. The technical design and implementation are mine. Evaluation experiments were conducted with the assistance of Chris Steven and Harmen de Weerd (University of Groningen, The Netherlands), and Peter van Rosmalen (Open University, The Netherlands); the interpretation of the results including meaningful comparisons are mine.

of dialogue contributions in term of effects on the information states of the dialogue participants. ISU models account for the creation of (shared) beliefs and mechanisms for their transfer, and have well-defined machinery for tracking, understanding and generation of natural human dialogue behaviour.

Cognitive modelling of human intelligent behaviour, as discussed in Chapter 3, enables deep understanding of complex processes related to human perception, comprehension, prediction, learning and decision making. The traditional information state specification can be extended with the representations of complex (multi-tasking) human multimodal behaviour.

The Cognitive Task Analysis (CTA) model incorporates expert knowledge and strategies obtained using common knowledge-eliciting techniques, requirements analysis and data from research literature, empirical evidence from real-life examples of expert and novice interactions, and data from small-scale controlled experiments. The CTA model enables to make detailed predictions of the effectiveness (success and quality) and efficiency (efforts) of the executed and practiced tasks and actions. It also supports the development of the skills of the virtual intelligent tutor which does not only enable to detect errors but also can explain why the learner's actions were incorrect.

Threaded cognition (Salvucci and Taatgen, 2008) and Instance-based Learning (IBL) (Gonzalez and Lebiere, 2005) models developed within the ACT-R cognitive architecture (Anderson, 2007) are used to design a cognitive agent that can respond and adapt to new situations, in particular to a communicative partner changing task goals and strategies. The agent is equipped with Theory of Mind skills (Premack and Woodruff, 1978) and is able to use its task knowledge not only to determine its own actions, but also to interpret the human partner's actions, and to adjust its behaviour to whom it interacts with. In this way, we expect to achieve flexible adaptive dialogue system behaviour in dynamic non-sequential interactions. The integrated cognitive agent does not only compute the most plausible task action(-s) given its understanding of the partner's actions and strategies, but also provides alternatives and plans possible outcomes, and knows why it selects a certain action and can explain why its choices lead to the specific outcome. This enables the agent to act as a cognitive tutor, supporting the development of the (meta)cognitive skills of a human learner. Finally, the agent can be built with rather limited real or simulated dialogue data: it is supplied with initial state-action templates encoding domain knowledge and the agent's preferences, and the agent further learns from the collected interactive experiences, see Section 4.4.

As ACT-R based computational cognitive models of Threaded cognition and IBL can be used to design cognitive agents that simulate task-related behaviour showing close to human decision-making performance. If such agents have Theory of Mind (ToM) skills they can exhibit metacognitive capabilities that are beneficial for better understanding and adequate modelling of adaptive and proactive task behaviour. They cannot yet deliver natural human-like interactive performance, but combining them with interactive agents based on advanced computational dialogue models opens new possibilities.

5.1 Dialogue Manager architecture

Inspired by the distinction that can be made between *task control* actions and *dialogue control* actions (Bunt, 1994), we explored these possibilities by integrating a cognitive task agent into the ISU-based dialogue manager.

In the dialogue system design community, involving both theorists and practitioners, a clean separation into two layers is observed. One layer deals with the task at hand, and the other with the communicative performance itself, see e.g. [Lemon et al., 2003]. To design task managers (agents), various approaches can be used. For instance, several approaches based on *hierarchical task analysis* have been proposed, see Section 3.5. The method has been used successfully to simulate human decision-making processes. In dialogue management, it has also been deployed in the form of *hierarchical task decomposition* and *expectation agenda generation* within the RavenClaw framework (Bohus and Rudnicky, 2003) and tested successfully in several systems. Examples include the use of a tree-of-handlers in the Agenda Communicator (Xu and Rudnicky, 2000), of activity trees in WITAS (Lemon et al., 2001), and of recipes in Collagen (Rich et al., 1998). However, models based on task hierarchies, agendas, recipes and trees are rather static and difficult to apply for non-linear (multi-branching) or non-sequential interactions, like multi-issue barganing dialogues.

A more flexible approach is the *plan-based* approach. For instance, in the TRIPS system (Allen et al., 2001) a Task Manager is implemented that relies on planning and plan recognition, and coordinates actions with a Conversational Manager. Plan construction and inference are activities that can easily get very complex, however, and become computationally intractable.

Multi-agent architectures have been proposed for adaptive and flexible human-computer interactions, e.g. in the JASPIS speech application (Turunen et al., 2005), in the Open Agent Architecture (Martin et al., 1999), and in Galaxy-II (Seneff et al., 1998).

An ISU-based approach to dialogue management has been used to handle multiple aspects ('dimensions') simultaneously separating task control acts and various classes of dialogue control acts (Keizer et al., 2011; Petukhova, 2011; Malchanau et al., 2015). The dialogue manager tracks updates in multiple dimensions of the participants' information states, as the effect of processing incoming multimodal dialogue acts, and generates multiple task control acts and dialogue control acts in response.

The above considerations have resulted in a Dialogue Manager consisting of multiple Agents corresponding currently to six ISO 24617-2 or DIT^{++} dimensions[1]: the Task Manager with the integrated Cognitive Task Agent (CTA) and Task Planner for task control, the Auto/Allo Feedback Agent, the Turn Manager, the Discourse Structuring Manager, the Contact Manager and the Social Obligations Manager.

In order to capture the dynamics related to frequently changing participants' interactive and strategic goals, we propose a flexible *adaptive* form of multidimensional dialogue management inspired by cognitive models of multitasking, learning and cognitive skills

[1]The set of Agents may in future be extended to include all nine ISO 24617-2 dimensions and possibly other additional dimensions.

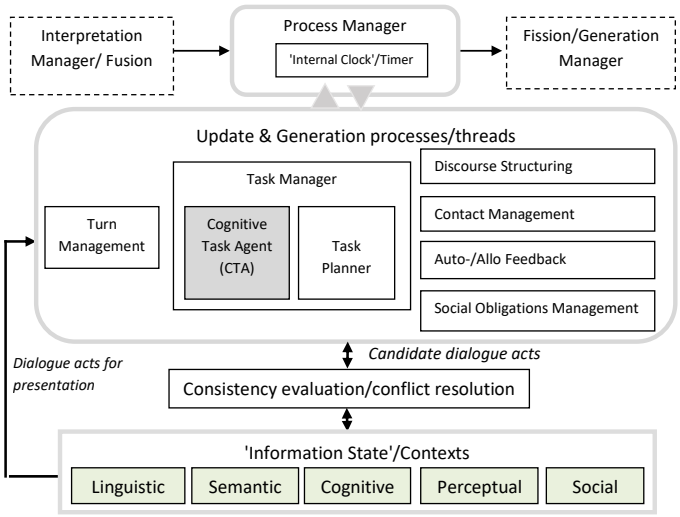

Figure 5.1: Cognitive Task Agent (grey box) incorporated into the Dialogue Manager architecture: dialogue acts are passed to the Dialogue Manager for context model update and next action(-s) generation; all processes are scheduled and executed by the Process Manager; Cognitive Task Agent maintains its own state (memory) internally, but transparently for the Task Manager.

transfer. To this end, we designed two Cognitive Task Agents and integrated them as part of an ISU-based multidimensional Dialogue Manager (DM).

The Dialogue Manager (DM) is designed as a set of processes ('threads') that receive data, update the information state, and generate output. Additionally, consistency checking and conflict resolution is performed to avoid that the context model is updated with inconsistent or conflicting information and incompatible dialogue acts are generated, see also Petukhova (2011). Figure 5.1 presents the overall DM architecture. First, data are received from the Fusion/Interpretation module. Next, the information state ('context model') is updated based on the received input. For this, the Process Manager dispatches incoming dialogue acts (based on their dimension) to the corresponding agents that in turn update relevant parts of the context model. Subsequently, the output based on the analysis of the information state is generated. The output presents the ordered list of dialogue acts which is sent to the Fission module, see next Section for complete dialogue system architecture.

5.2 Multimodal information state

According to the ISU approach, dialogue behaviour, when understood by a dialogue participant, evokes certain changes in the participants' information state. Since we deal with several different interactive, task-related and tutoring aspects, an articulate context model should contain all the information considered relevant for interpreting such rich multimodal dialogue behaviour in order to enable the system to generate an adequate reaction playing the role of an Observer, a Mirror, an Experiencer and that of a Tutor, see Section 3.2. As

an Observer, the system tracks natural multimodal argumentative and negotiation behaviour, and tries to understand it as well as possible. In the Mirroring mode, the systems replays the observed and analysed behaviour. As an Experiencer, the system interacts with the human learner as an active full-fledged dialogue participant, either as Negotiator or a Debater. As a Tutor, the system's task is to provide tutoring interventions in the form of the formative real-time "in-action" feedback. All system's roles require reliable tracking and understanding of rather complex multimodal behaviour related to three key aspects:

1. *presentational* aspects associated with credible debate and negotiation performance when applying appropriate multimodal rhetoric devices comprising a range of linguistic, paralinguistic and non-verbal behavioral properties;

2. *interactional* aspects that are important for any successful and pleasant interaction such as managing turns and time, ensuring contact and compliance with social norms and conventions; and

3. aspects related to the appropriate *negotiation strategies* and convincing *argumentation*.

Table 5.1 summarises what inappropriate behavioural patterns the system is able to 'hear' and 'see' when involved in debate and negotiation interactions.

A participant's multimodal dialogue behaviour is analysed in terms of dialogue acts, defined in accordance with the ISO 24617-2 annotation standard, assigned to multimodal segments that have certain meaning in terms of communicative functions in one or more dimensions. Several sensors capture the speech signal and noticeable motions. Speech signals are recorded from multiple sources, such as wearable microphones, headsets for each dialogue participant, and an all-around microphone placed between participants. The Kaldi-based Automatic Speech Recognition (ASR) system is used to compute the 1st best word sequence (Povey, 2011). Prosodic properties related to voice quality, fluency, stress and intonation of speech are computed using the PRAAT tool [Boersma and Weenink, 2009]. To track visible movement, depth sensing devices for full-body tracking such as Microsoft Kinect v2 were used. Further modern sensors enabling fine grained tracking of body movement and facial expressions (Intel®RealSense™, eye-trackers like Tobii Glasses) and various biometrical signals (Blood Volume Pulse and NeXus EXG sensors), can be considered for extensions. All inputs are time aligned and are temporally synchronised to represent primary behavioural data which contains indexed verbal and non-verbal elements. Subsequently, these modality-specific elements (often representing low-level surface features, e.g. numerical features with tracking information or tokenised speech transcription) are fused into a representation of user's actions, e.g. dialogue acts. Figure 5.2 provides a typed feature representation of a *dialogue act* assigned to a *functional segment* - a stretch of multimodal behaviour produced by a *sender* which has a (*qualified*) *communicative function* in one or multiple *dimensions* and is linked through various *dependence* and *rhetorical relations* to the segments or dialogue acts back in the dialogue history. A dialogue act has a certain *semantic content* which is often domain-specific and is computed from features of given and related functional segments.

Behavioural aspect		Interactive setting	
		Debate	Negotiation
Presentation	speech fluency	$>$ 7 silent pauses that are $>$ 200*ms* per speaking turn	
	speech volume	too loud ($>$60 decibel); too soft ($<$30 decibel)	
	hand and arm position	arms crossed; hands invisible (e.g. in the pockets, behind the back too much gesticulation: $>$ 70% of unclassified gesture events per turn	
	posture	slouching	
Interaction	turn management	abrupt and frequent interruption of other speaker (overlapping speech)	
	time management	long turns: $>$2 minutes	long turns: $>$4 dialogue acts
	social obligations	absence of returnGreetings, returnGoodBye, (accept)Thanking; (accept)Apologies	
Content	structure	missing/unmarked justifications/evidence: no discourse marker detected	repetitive rejections: DeclineOffer sequences of length $>$2
	semantics	low relevance of arguments: $>$ 50% out-of-vocabulary tokens	non-cooperative negotiation moves: no BargainDown or Concessions
		low clarity of the proposed arguments or negotiation moves: high number of syntactic/semantic chunks: $>$ 24 syntactic constituents per argument or $>$ 10 per negotiation turn high number of referring expressions: $>$ 7 referring expressions per ADU or $>$ 5 per negotiation turn	
	outcome	low acceptability of the proposed argument: no agreements; 100% difference in participants' final debate states	dead-lock situation or termination: Withdraw, Exit or BlockAgreement moves detected

Table 5.1: Overview of the inappropriate behaviour distinctive for debate and negotiation settings, extension of Petukhova et al. (2017) and van Helvert et al. (2015). Where possible, measurable indicators are provided.

$$
\left[\begin{array}{l}
DialogueAct(DA): \left[\begin{array}{l}
dimension(D): \langle dim \rangle \\
comm\_function(CF): \langle cf \rangle \\
sem\_content(SC): \langle content \rangle \\
sender/speaker: \langle participant \rangle \\
addressee: \langle participant \rangle \\
functional\_dependence: \left[\ antecedent: \{\langle DA \rangle\}\ \right] \\
feedback\_dependence: \left[\ antecedent: \{\langle FS \rangle\}\ \right] \\
rhetorical\_relation: \left[\ antecedent: \{\langle DA \rangle\}\ \right] \\
FunctionalSegment(FS): \left[\begin{array}{l}
start: \langle token_{15}|1260 \rangle \\
end: \langle token_{24}|5640 \rangle \\
verbatim: \langle token_{index}|time\_points = \text{`}token1', \ldots \rangle \\
prosody: \langle duration, pitch, energy, \ldots \rangle \\
nonverbal: \langle \left[\begin{array}{l}
head\langle element_{index}|time\_points = expression1, \ldots \rangle \\
hands\langle element_{index}|time\_points = expression1, \ldots \rangle \\
face\langle element_{index}|time\_points = expression1, \ldots \rangle \\
posture\langle element_{index}|time\_points = expression1, \ldots \rangle
\end{array}\right] \rangle \\
sender: \langle participant \rangle
\end{array}\right]
\end{array}\right]
\end{array}\right]
$$

Figure 5.2: Example of feature structure representation of a dialogue act.

Complexities of natural multimodal human dialogue are handled by analysing dialogue behaviour as having communicative functions in several dimensions, as discussed in Section 2.6. An articulate dialogue model and context model have been proposed by Bunt (1999). The context model has five components:

1. **Linguistic Context (LC)**: information about dialogue acts (1) produced up to this point ('dialogue history'); (2) most recently produced dialogue act ('latest state'); and (3) planned future contributions ('dialogue future' or 'planned state'). Participants' beliefs concerning interpreted behaviour and speaking roles are also modelled here.

2. **Semantic Context (SemC)**: information about the task that includes representation of (1) task progress and success; (2) speaker's beliefs about the domain ('domain knowledge' obtained from Discourse Model); (3) speaker's beliefs about the dialogue partner's semantic context.

3. **Cognitive Context (CC)**: information about (1) the current processing state of the speaker; (2) assumptions and expectations about the partner's cognitive context; (3) estimation of time needed for processing of the current contribution.

4. **Perceptual/Physical Context (PC)**: information about the perceptible aspects of the communication process and the task/domain such as speaker's presence and readiness to continue the dialogue and assumptions about partner's perceptual/physical context.

5. **Social Context (SocC)**: information about current speaker's (1) interactive pressures and (2) reactive pressures, and assumptions and expectations about partner's social context.

Each of these five components contains the representation of three parts: (1) the speaker's beliefs about the task, about the processing of previous utterances, or about certain aspects of the interactive situation; (2) the addressee's beliefs of the same kind, according to the speaker; and (3) the beliefs of the same kind which the speaker assumes to be shared (or

$$
\left[\begin{array}{l}
\mathit{LingContext}: \left[\begin{array}{l}
\mathit{speaker}: \left[\begin{array}{l}
\mathit{dialogue\_history}: \left\{ \langle \mathit{previous\_DA}\ \boxed{\text{Fig. 5.2}}\ \rangle \right\} \\
\mathit{latest\_state}: \left[\begin{array}{l} \mathit{DA}: \boxed{\text{Fig. 5.2}} \\ \mathit{state} = \mathit{opening}|\mathit{body}|\mathit{closing} \end{array}\right] \\
\mathit{dialogue\_future}: \mathit{plan}: \left[\begin{array}{l} \mathit{candidates}: \langle \mathit{list\_DAs} \rangle \\ \mathit{order}: \langle \mathit{ordered\_list\_DAs} \rangle \end{array}\right]
\end{array}\right] \\
\mathit{partner}: \langle \mathit{partner\_linguistic\_context} \rangle\ (\mathit{according\ to\ speaker}) \\
\mathit{shared}: \langle \mathit{shared\_linguistic\_context} \rangle
\end{array}\right] \\
\mathit{SemContext}: \left[\begin{array}{l}
\mathit{speaker\_task\_model}: \langle \mathit{beliefs} \rangle \\
\mathit{partner\_task\_model}: \langle \mathit{beliefs} \rangle\ (\mathit{according\ to\ speaker}) \\
\mathit{shared\_task\_model}: \langle \mathit{mutual\_beliefs} \rangle
\end{array}\right] \\
\mathit{CogContext}: \left[\begin{array}{l}
\mathit{speaker\_own\_proc\_state}: \left[\begin{array}{l}
\mathit{proc\_problem}: \mathit{yes}|\mathit{no} \\
\mathit{problem\_input}: \mathit{FS} \\
\mathit{time\_need}: \mathit{negligible}|\mathit{small}|\mathit{substantial}
\end{array}\right] \\
\mathit{partner\_proc\_state}: \langle \mathit{partner\_cognitive\_context} \rangle (\mathit{according\ to\ speaker}) \\
\mathit{shared}: \langle \mathit{shared\_cognitive\_context} \rangle
\end{array}\right] \\
\mathit{PercContext}: \left[\begin{array}{l}
\mathit{speaker}: \left[\begin{array}{l}
\mathit{own\_presence}: \mathit{positive}|\mathit{negative} \\
\mathit{own\_readiness}: \mathit{positive}|\mathit{negative}
\end{array}\right] \\
\mathit{partner}: \langle \mathit{partner\_perceptual\_context} \rangle (\mathit{according\ to\ speaker}) \\
\mathit{shared}: \langle \mathit{shared\_perceptual\_context} \rangle
\end{array}\right] \\
\mathit{SocContext}: \left[\begin{array}{l}
\mathit{speaker}: \left[\begin{array}{l}
\mathit{interactive\_pressure}: \mathit{none}|\mathit{greet}|\mathit{apology}|\mathit{thanking}|\ldots \\
\mathit{reactive\_pressure}: \{\langle \mathit{dialogue\_acts} \rangle\}
\end{array}\right] \\
\mathit{partner}: \langle \mathit{partner\_social\_context} \rangle (\mathit{according\ to\ speaker}) \\
\mathit{shared}: \langle \mathit{shared\_social\_context} \rangle
\end{array}\right]
\end{array}\right]
$$

Figure 5.3: Feature structure representation of the context model. Updated of Malchanau et al., 2015.

'grounded') with the addressee. Figure 5.3 shows the context model with its component structure.

A communicative function specifies how an understanding dialogue participant's context model is updated, where the dimension (semantic content type) determines which parts of the context model are updated. The information state of an addressee of a dialogue act changes when he understands the speaker's behaviour. For instance, the Linguistic Context is updated when dealing with presentational aspects and some interactional aspects, such as turn management; in the Cognitive Context participants' processing states are modelled, as well as aspects related to time and own communication management (e.g. error in speech production). The Semantic Context contains representations of task-related actions, e.g. negotiation actions or debate arguments (system as Negotiator or Debater respectively), and/or the system's tutoring goals and expectations on a trainee's learning progress (system as a Tutor), and/or all of these if the system is in observing and mirroring mode.

5.3 Dialogue acts update semantics

Semantically, dialogue acts correspond to update operations on the information states of the dialogue participants. They have two main components: (1) the *communicative function*, that specifies how to update an information state, e.g. Inform, Question, and Request, and (2) the *semantic content*, i.e. the objects, events, situations, relations, properties, etc. involved in the update, see [Bunt, 2000]. An utterance, when understood by a dialogue participant as a dialogue act with a certain communicative function and semantic content, evokes certain changes in the participant's information state (context model). Dialogue acts are formally defined as operators that have certain update effects on the speaker's and addressees' context models and are characterised in terms of *preconditions*, *effects* and a *body* that describes the means by which effects are achieved (Cohen and Perrault, 1979). To describe the intended update effects of an action a number of formal concepts - semantic primitives - are used that specify an agent's beliefs, goals, and commitments. A set of semantic primitives is defined in [Petukhova, 2011]. Bunt (2014) provides a detailed specification of the update semantics of dialogue acts. For instance, the primitive *Bel* expresses the possession of information and the *KnowVal* primitive serves to represent the availability of information. For example, *A* believing that *B* has certain preferences for the 'scope' issue is represented as $Bel(A, KnowVal(B, prefer(ISSUE = 1; ?VALUE)))$.[2] The primitive *Want* is used to capture a participant's goal to achieve a certain situation. Thus, *A*'s goal to obtain information about *B*'s negotiation preference can be represented as $Want(A, KnowVal(A, prefer(ISSUE = 1; ?VALUE)))$. Consider, for an example, the update semantics of a Suggestion act:

Update definition (formalised)	Description
$Want(S, ConsidDo(A, \alpha, A, C_\alpha))$	Speaker *S* wants that the addressee *A* considers to perform the action α, in the manner or with the frequency described in the semantic content, i.e. under certain conditions C_α;
$Bel(S, Interest(A, \alpha))$	S believes that α is promising and of interest to *A*, which is specified as part of the semantic content;
$Assume(S, CanDo(\{A, S\}, \alpha))$	*S* assumes that *A* (possibly together with S) is able to perform α in the manner or with the frequency described.

5.3.1 Debate semantics

In a debate, participants argue *for* or *against* a certain motion or issue. Thus, they mostly exchange arguments consisting of Inform acts related by means of various rhetorical relations, see [Petukhova et al., 2015c] and [Petukhova et al., 2017b]. Frequently, the main claim, i.e. a Statement, is supported by either a Reason or Evidence, and is wrapped up by a Conclusion. For example:

(12) $D1_{21}$[3]: Past anti-smoking campaigns were useless [*Inform*]
$D1_{22}$: I haven't actually seen any of those implemented [*Inform Motivate* $D1_{21}$]

[2] Additionally, the strength of *A*'s beliefs is represented by the parameter σ, which can have the values 'firm' and 'weak', or numerical values expressed, for example, by confidence scores computed elsewhere.

[3] Here and henceforth D*k* stands for Debater *k*; the subscript is the index of the identified dialogue act.

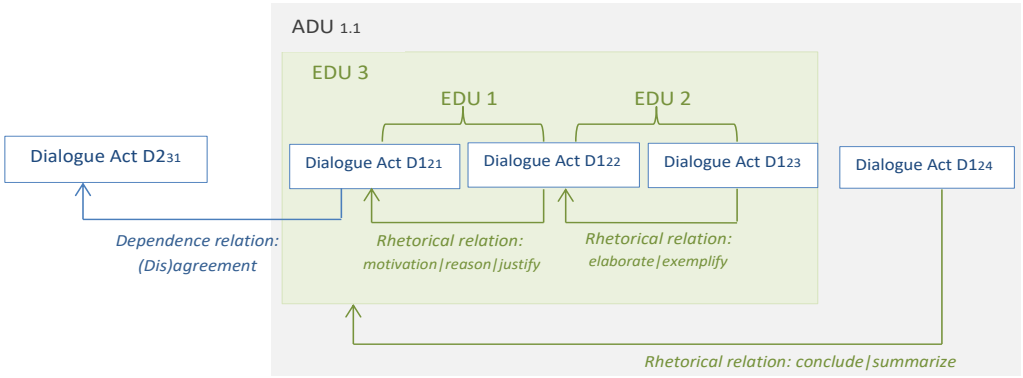

Figure 5.4: Identification of Argumentative Discourse Units (ADU).

D1$_{23}$: I have personally walked into a store and seen a fourteen years old buying a pack of cigarettes [*Inform Exemplify* D1$_{21}$]
D1$_{24}$arg1.1[4]: Many cases of civil disobedience make this campaign look nice only on paper [*Inform Conclude* D1$_{21}$]

Thus, the most frequent updates are concerned with the information exchange in form of arguments, e.g. believing that p, $Bel(S,p)$, and having the goal that the addressee also believes that p, $Want(S,Bel(A,p))$. An argument content p can be rather complex, specifying events, event participants and various types of semantic and discourse relations between events and between participants.

In our application scenario and according to the model of the hierarchical debate training tasks (see Figure 4.4), the debate semantics is specified to facilitate (1) the recognition of the supportive and rebuttal arguments presented by the debater to achieve minimal debate goals, and (2) the generation of the system in-action feedback on argument structure, quality and delivery aspects of the observed debater performance.

In the first task, the context update operation will correspond to the Inform act with the semantic content specified in the main conclusion of the identified argument. If the conclusion is not available (or not detected), the semantic content will be computed from the main claim. For an example above in 12, $Bel(D_1,D1_{1.1})$ and $Wants(D_1,Bel(D_2,D1_1.1)$.

The proposed argument identification process is illustrated in Figure 5.4, see also Petukhova et al. (2016) for details on processing and segmentation. The process starts with segmenting a debater's turn into functional segments each of them having one or more communicative functions according to the ISO 24617-2 *dialogue act* annotation standard. Subsequently, we propose to identify *rhetorical* (or discourse) *relations* between dialogue acts that are mostly Informs and cluster them into *Elementary Discourse Units* (EDU), and successively into *Argumentative Discourse Units*. For this, the PDTB discourse parser was used[5] and a discourse relation classifier was trained [Petukhova et al., 2017b]. The ADU's

[4]Here and henceforth *x.y* is the index assigned to the conclusion of an Argumentative Discourse Unit (ADU), where *x* indicates the debater index and *y* stands for the index of an ADU conclusion.

[5]Visit `https://wing.comp.nus.edu.sg/~linzihen/parser/` for detailed informa-

main statement (claim) can then be extracted, which is either the opening Inform or the closing Conclusion or Re-statement. These propositions are then linguistically processed using state-of-the-art parsers of various types, e.g. a syntactic parser and (shallow) semantic parsers. One of the tools that incorporates many of the required existing up-to-date semantic analysers is Boxer[6]. It takes as input CCG (Combinatory Categorial Grammar) derivations and produces DRSs (Discourse Representation Structures) with an option to convert to SDRSs (Segmented Discourse Representation Structures) to capture rhetorical relations as illustrated in Figure 5.5. In many cases the identification of attack/support links requires an additional step, since our analysis showed most of them are expressed by explicit or implicit (dis-)agreements. Arguments represented by their main propositions (either claims or conclusions) and support/attack links between them are semantically modelled as part of the debaters' information states (see Section 5.6).

The second task is concerned with the system interpretation of the appropriateness of the debaters' presentation and interactive behaviour and the generation of corrective feedback. Criteria on which basis the Debate Coach makes his decisions are summarised in Table 5.2. For example, the following behaviour is recognised[7]:

```
dialogActId: da_p1_2
communicativeFunction: inform
start: 1.67
end: 5.97
sender: p1
addressee: p2
verbatim: Smoking should not be banned in all public places
speakingVolume: High
gesture: ARMSCROSSED
```

The system believes to have interpreted the participant P1 dialogue contribution as having certain semantic content (representation of verbal component, speakingVolume:HIGH and gesture:ARMSCROSSED, which P1 believes are correct) and communicative function (Inform). Using the knowledge available to the system, e.g. in a database with prosodic properties and visible body movements that are inappropriate in a debate situation (see Table 5.1 for details), the system's task is to inform the addressee about his presentational failures. More formally, $Bel(S, \neg appropriate(volume(fs_1) = high))$; $Bel(S, \neg appropriate(gesture(fs_1) = ARMSCROSSED))$; $Want(S, Bel(P_1, \neg appropriate(volume(fs_1) = high))$; and $Want(S, Bel(P_1, \neg appropriate(gesture(fs_1) = ARMSCROSSED))$.

5.3.2 Negotiation semantics

Negotiations are commonly analysed in terms of certain actions, such as offers, counter-offers, and concessions, see [Watkins, 2003], [Hindriks et al., 2007]. We considered two

tion and download.

[6] http://svn.ask.it.usyd.edu.au/trac/candc/wiki/boxer

[7] NOTE: for the simplicity we provide here dialogue act representation in the JSON format

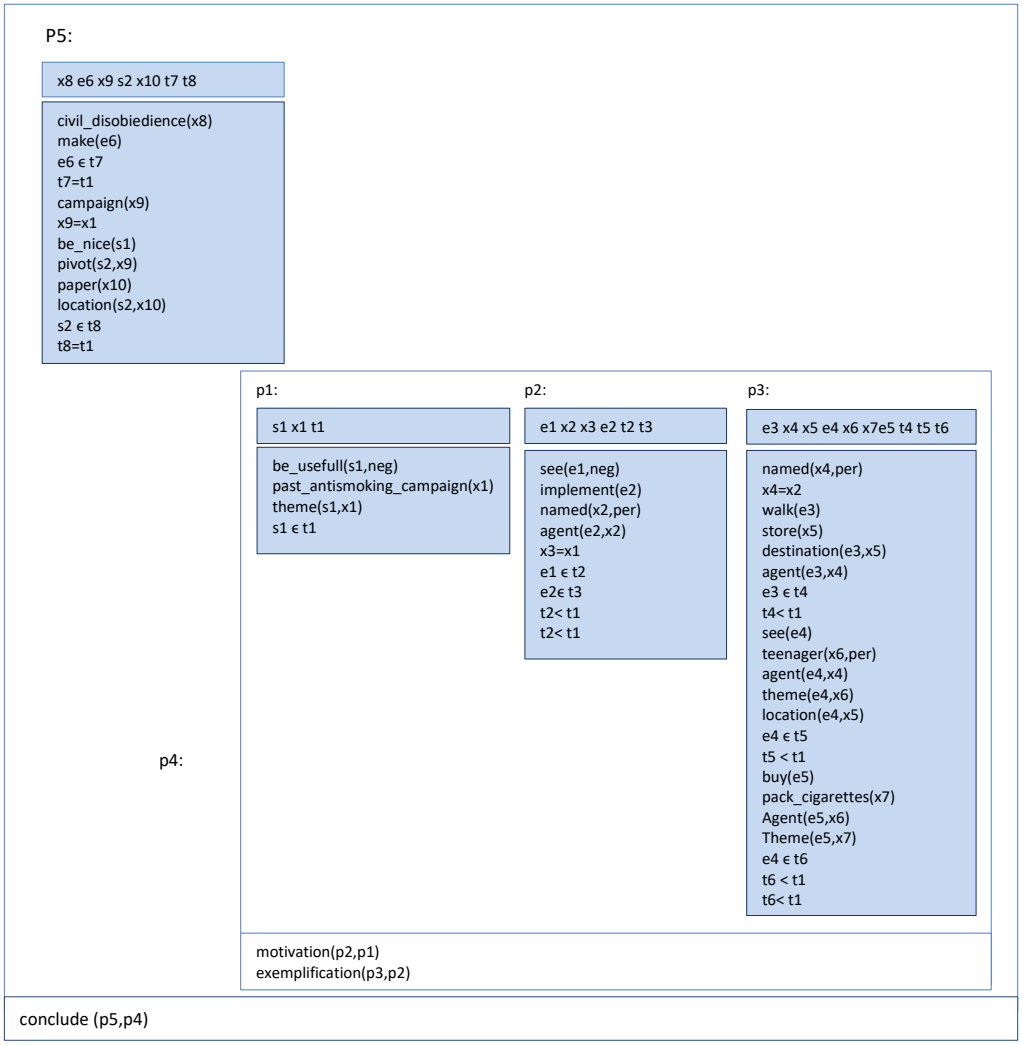

Figure 5.5: Example of SDRS representation of the identified ADU presented in (12).

possible ways of using such actions, also referred to as 'negotiation moves', to compute the update semantics in negotiation dialogues. One is to treat negotiation moves as task-specific dialogue acts. Due to its domain-independent character, the ISO 24617-2 standard does not define any communicative functions that are specific for a particular kind of task or domain, but the standard invites the addition of such functions, and includes guidelines for how to do so. For example, a negotiation-specific kind of $Offer_N$ function should be introduced for the expression of commitments concerning a negotiation value.[8] Another possibility is to use negotiation moves as the semantic content of general-purpose dialogue acts. For example, a negotiator's statements concerning his preference to a certain option can be represented as $Inform(A,B,\Diamond offer(X;Y))$.

It has been observed that in negotiations, participants exchange offers expressing different levels of commitment with respect to the targeted negotiated outcome (Raiffa et al., 2002). We distinguished five levels of commitment:

1. zero commitment for offer elicitations and preference information requests, e.g. a Set Question of participant A addressed to B with the goal to elicit B's preference concerning the smoking ban scope, e.g. 'Where do you think we should ban smoking?', can be represented as $SetQuestion(A,B,offer(ISSUE = 1;?VALUE))$;

2. the lowest non-zero level of commitment for informing about preferences, abilities and necessities, e.g. A's goal to inform B about his negotiation preferences can be represented as $Want(A,KnowVal(B,Bel(A,offer(ISSUE = 1;VALUE = 1C))$;

3. an interest and consideration to offer a certain value, e.g. Suggestion expressing considerations to offer certain values, and assumptions about the opponent's abilities and interests to offer the same, i.e. $ConsidDo(A,offer(X;Y))$; $Bel(A,CanDo(B,offer(X;Y))$); $Bel(A,Interest(B,offer(X;Y)))$;

4. weak (tentative) or conditional commitment to offer a certain value, e.g. all offers and responses to them at all negotiation stages except for the final phase are modelled as weak commitments, e.g. $WBel(A,CommitDo(A,offer(X;Y)))$, indicating that they are *tentative*, and can eventually be strengthened or cancelled, see also Section 5.6; and

5. strong (final) commitment to offer a certain value, e.g. Offer acts are observed, expressing commitments to offer (or not to offer) a certain value, e.g. $CommitDo(A,offer(X;Y))$ and $CommitRefrain(A,offer(X;Y))$

8 basic negotiation moves are defined: offer, counter-offer, exchange, concession, bargain-in, bargain-down, deal and withdraw.

Negotiators often communicate their cooperativity by using modal utterances expressing preference and ability. Non-cooperative behaviour, by contrast, may be articulated by expressing inability and dislike. Modality expressions are mainly observed in *Inform* and *Answer* acts.

[8] Negotiation 'Offers' may have a more domain-specific name, e.g. *Bid* for selling-buying bargaining.

The <negotiationSemantics> element has been added to DiAML to represent the semantic content of a dialogue act. A shallow negotiation semantics is defined in terms of <negotiationMove> with attributes defined for different types of such moves. For example:

```
<dialogueAct xml:id="dap1TSK38" sender="#p1"
             addressee="#p2"
   dimension="task" communicativeFunction="inform"
   target="#fsp1TSKCV38">
     <negotiationSemantics>
       <negotiationMove type="counterOffer"/>
     </negotiationSemantics>
   <rhetoricalLink rhetoAntecedent="#dap2TSK37"
                   rhetoRel="substitution"/>
 </dialogueAct>
```

Additionally, dependent on annotation goals, approach, granularity and type of semantic processing, <negotiationSemantics> elements can be extended with elements, based on <Arg> type, for negotiated issues, values and logical operators between arguments. Modal relations can be represented by <modalLink> linking the holder (e.g. speaker) and target (semantic content) with values describing the speaker's attitudes to the necessity or probability of the events, and the speaker's abilities. The full proposed DiAML representation of utterance *P1: I prefer all outdoor smoking allowed* produced by the sender P1 addressed to P2 is a task-related Inform act with the semantic content $\Box offer(1a)$ is as follows:

```
<dialogueAct xml:id="da1" sender="#p1" addressee="#p2"
  dimension="task" communicativeFunction="inform"
  target="#fsp1TSKCV38" qualifier="certain">
    <negotiationSemantics>
      <negotiationMove xml:id="nm1" type="offer">
        <arg>
          <value>1a</value>
        </arg>
        <modalLink holder="#p1" target="#nm1"
                   modalRel="preference"/>
      </negotiationMove>
    </negotiationSemantics>
</dialogueAct>
```

5.4 Cognitive Task Agents

The Cognitive Task Agent (CTA) operates on a structured dynamic Semantic Context as described above, identifies the partner's task-related goals, and uses this information to compute its next action(-s). Two Cognitive Task Agents are designed - Debate Coach Agent and Negotiation Agent.

5.4.1 Debate Coach Agent

The Debate Coach Agent (DCA) gets interpretations of multimodal natural debate behaviour and provides tutoring interventions on the trainee performance on aspects related to argument content and perceptual and physical presentational debate aspects. The articulate context model specified above supports adequate understanding of debaters' rich multimodal dialogue behaviour and enables the tutoring system to generate an adequate reaction related to several different tutoring aspects.

In the current implementation, the DCA's tasks are concerned with formative ('in-action') feedback generation on presentation aspects of the debater performance and with summative ('about-action') feedback on the success or failure of the trainee in achieving his goals. The DCA operates on the basis of the cognitive task analysis model as a part of the hierarchical whole-task model presented in Section 4.4.

As for argument delivery, five core aspects are considered: Audibility, Engagement, Conviction, Authority and Likability (AECAL). Although it is often difficult to define clear properties of good debate or public speaking, there are certain linguistic, prosodic and body language features that correlate with human judgments of such behaviour. Debaters make use of these features which enables explaining perceptive regularities and to formulate argumentation strategies that trainees may follow to deliver convincing debate performance, and that can be used for its assessment. Table 5.2 summarises previous findings on correlations observed between linguistic, acoustic, prosodic properties of speech and visible body movement properties, and human judgments of a 'good rhetoric' linked to AECAL aspects.[9] The presented correlates are not only powerful communicative tools used by skilled debaters to persuade their audience, but they also influence discourse processing to a great extent (see e.g. [Dahan et al., 2002, Watson et al., 2008, Repp and Drenhaus, 2015]).

The DCA currently acts based on the criteria defined in Table 5.1. The criteria from Table 5.2 can be used for further extensions when more robust recognition and interpretation of these behaviours will be enabled.

[9]The presented matrix is rather simplified. In reality the mapping is not 1:1 and cross-factor dependencies exist.

Delivery aspects	Performance strategy	Correlates		
		linguistic	acoustic-prosodic	visible body movements
Audibility	Adequate voice volume Appropriate argumentation pace	- number of tokens per second	perceived as normal (60-54 dB) number of syllables per second	- number of beats per second
Engagement	Expressiveness	> repetitions (List of Three) [Beard, 2002] > personal pronouns density [Rosenberg and Hirschberg, 2009] < information density and redundancy [Nir, 1988, Touati, 2009]	variations in pitch range [Touati, 1993] > standard deviation in pitch [Rosenberg and Hirschberg, 2009, Touati, 1993]	open gestures (palm) appropriate gesticulation
Conviction	Clear articulation and fluency Adequate prominence and focus, topic-comment structuring	no disfluences and hesitations [Tuppen, 1974, Braga and Marques, 2004] no false start [Rosenberg and Hirschberg, 2009] topicalisation, passivisation, it- and wh-cleft discourse structuring or meta-discoursive acts [Nir, 1988, Touati, 2009]	fraction of voiced/unvoiced frames frequent voice breaks > pitch range; > mean pitch; > intensity [Hirschberg, 2002, Rosenberg and Hirschberg, 2009] [Pejčić, 2014, Braga and Marques, 2004, Touati, 2009] emphatic accents [Novák-Tót et al., 2017]	hand & arm position posture (e.g. no sloutching) adequate beat gestures iconic & metaphoric gestures
Authority	Adequate grouping & phrasing	clear syntactic structures,phrasing, chunking [Touati, 2009]	slowing down speech rate [Strangert, 1991, Touati, 2009] pausing [Wichmann, 2002, Strangert, 2005, Touati, 2009]	confident posture
Likability	Express respect and friendliness	sentiment vocabulary, e.g. affect dictionary [Whissell, 2009] sentiment shifters; offensive language use [Warner and Hirschberg, 2012]	pitch register	eye contact, smiling

Table 5.2: Properties of persuasive public speech (as judged by humans) and their lexico-syntactic, acoustic-prosodic and motion correlates as observed in previous empirical studies. Adopted from Petukhova et al. 2017

In more detail, when the system recognises that debater D1 occupies the speaker role (i.e. has a turn) and interprets his/her behaviour as D1 speaking too loud and/or performing an inappropriate body movement, it should react by either informing the addressee of his infelicitous use of voice and body, or propose how this behaviour can be corrected. At the same time the system does not want to take the turn over, but rather communicate its messages in a non-intrusive manner. Thus, system responses are generated visually using colors (red meaning something wrong happened, green - participant's performance is according to expectations, thus providing positive feedback) and pictures depicting correct body position, plus a verbal message, e.g. *'Reset your posture'*. The context model is updated accordingly, as shown in Table 5.8 where also full updates and beliefs transfer mechanisms are specified illustrating how they lead to the generation of specific tutoring interventions.

To generate the summative feedback on the debaters' overall task success, it is necessary, first, to detect the arguments (or Argumentative Discourse Units, see method proposed in Petukhova et al., 2016) and indicate what claims and relations do occur; and then to establish how these relations are verified during a debate session. The DCA computes the strength and sustainability of arguments. The DCA beliefs are concerned with computing supporting and attacking arguments to make a prediction about the debate outcome. The debate outcome predicted by the system is compared with the minimal goals assigned to the debaters and the differences (if any) are used for the summative feedback generation. For this, the ADU's main statement is identified and attack/support links are computed as discussed in Section 5.3. Debaters's information states are updated. Further, the tracking of created shared beliefs and beliefs transfer is performed to compute the debate final information states which were used to assess debaters' successes and failures in achieving the goals assigned to them, see Section 5.6 for details.

5.4.2 Negotiation Agent

The Negotiation Agent (NA) is designed to be integrated as a part of the Virtual Negotiation Coach application - an intelligent tutoring system to train metacognitive skills in negotiation setting.

For many existing human-computer negotiation systems, interactions are typically modelled as a sequence of competitive offers where partners claim a bigger share for themselves. Valuable work has been done on well-structured negotiations where few parties interact with fixed interests and alternatives, see e.g. [Traum et al., 2008], [Georgila and Traum, 2011], [Guhe and Lascarides, 2014], [Efstathiou and Lemon, 2015]. In many real-life negotiations, parties negotiate not over one but over multiple issues, see e.g. [Cadilhac et al., 2013], where they have interests in reaching agreements about several issues, and their preferences concerning these issues are not completely identical (Raiffa et al., 2002). Negotiators may have partially competitive and partially cooperative goals, and may make trade-offs across issues in order for both sides to be satisfied with the outcome. Parties can delay making a complete agreement on the first discussed issue, e.g. they postpone making an agreement or make a partial agreement, until an agreement is reached on the second one. They can revise their past offers, accept or decline any standing offer, make counter-offers, etc. We con-

sider such complex strategic negotiations as multi-issue integrative bargaining dialogues, see Petukhova et al. (2016 and 2017). We aim at modelling these interactions with the main goal to train metacognitive skills. Comparable work has been performed on modelling so-called semi-cooperative multi-issue bargaining dialogues, see (Lewis et al., 2017), who proposed an approach to end-to-end training of negotiation agents using a dataset of human-human negotiation dialogues, and applying reinforcement learning. Their study presents a new form of planning ahead where possible complete dialogue continuations are simulated - *dialogue rollout*. Our approach also allows to compute the best alternative move at each negotiation stage and plan ahead the complete negotiation. We compute about 420 outcomes per scenario, for 9 scenarios in total, each featuring different preference participant profiles. Additionally, for tutoring purposes the model provides an explanation for all alternative choices and how they lead to what outcomes. The two approaches differ with respect to the amount of data/resources used (our 50 vs 5808 dialogues); scenario complexity (4 issues, 16 values and 9 different preference profiles in our scenario vs 3 types of items and 6 objects in Lewis et al., 2017); and modalities modelled (multimodal vs typed conversations). In our study, we explicitly model various negotiation strategies, while in Lewis et al. (2007), evidence of such strategies is observed, e.g. compromising or deceiving, and are implicitly learned but not considered by design.

Negotiation strategies

The specific negotiation setting considered here is a multi-issue bargaining scenario. Traum et al. (2008), who also consider a multi-issue bargaining setting, but viewed as a multi-party problem-solving task, define strategies as objectives rather than the orientations that lead to them. They distinguish seven different strategies: find issue, avoid, attack, negotiate, advocate, success and failure. Other researchers define negotiation strategies closely related to conflict management styles, i.e. the overall approach for conducting a negotiation. Five main strategies are observed: competing (adversarial), collaborating, compromising, avoiding (passive aggressive), and accommodating (submissive), see [Raiffa et al., 2002, Tinsley et al., 2002]. As in integrative negotiation, where the negotiators strive to achieve a delicate balance between cooperation and competition (Lax and Sebenius, 1992), we define two basic negotiation strategies: cooperative and non-cooperative.

Cooperative negotiators share information about their preferences with their opponents, are engaged in problem-solving behaviours and attempt to find mutually beneficial agreements (De Dreu et al., 2000). A cooperative negotiator prefers the options that have the highest collective value. If not enough information is available to make this determination, a cooperative negotiator will elicit this information from his opponent. A cooperative negotiator will not engage in positional bargaining[10] tactics, instead, he will attempt to find issues where a trade-off is possible.

Non-cooperative negotiators prefer to withhold their preferences in fear of weakening their power by sharing too much, or they may not reveal true preferences deceiving and misleading the partner. These negotiators focus on asserting their own preferred positions

[10]Positional bargaining involves holding on to a fixed set of preferences regardless of the interests of others.

rather than exploring the space of possible agreements (Fisher and Ury, 1981). A negotiator agent using this strategy will rarely ask an opponent for preferences, and will often ignore a partner's interests and requests for information. Instead, a non-cooperative negotiator will find his own ideal offer, state it, and insist upon it in the hope of making the opponent concede. He will threaten to end the negotiation or will make very small concessions. The non-cooperative negotiator will accept an offer only if he can gain from it.

We also model a *neutral* (or cautious) strategy. Neutral actions describe behaviours that are not indicative of either strategy above.

The Agent adjusts its strategy according to the perceived level of the opponent's co-operativeness. The Agent starts neutrally, requesting the partner's preferences. If the Agent believes the opponent is behaving cooperatively, it will react with a cooperative negotiation move. For instance, it will reveal its preferences when asked for, it will accept the opponent's offers, and propose concessions or cross-issues trade-offs. It will use modality triggers of liking and ability. If the Agent experiences the opponent as non-cooperative, it will switch to non-cooperative mode. It will stick to its preferences and insist on acceptance by the opponent. It will repeatedly reject the opponent's offers using modal expressions of inability, dislike and necessity. It will rarely make concessions. It will threaten to withdraw reached agreements and/or terminate negotiation. Such meta-strategies for strategy adjustment are observed in human negotiation and coordination games, see [Kelley and Stahelski, 1970], [Smith et al., 1982]. We explain in some detail how this is implemented.

To sum up, our approach is based on the cognitive negotiation model of integrative multi-issue bargaining, which incorporates potentially different beliefs and preferences of negotiation partners, learns to reason about these beliefs and preferences, and accounts for changes in participants' goals and strategies.

Instance design: creation, activation and retrieval

The Agent's negotiation moves and their arguments are encoded as 'instances' represented as a set of slot-value pairs corresponding to the Agent's preference profile. Information encoded in an instance concerns beliefs about Agent's and partner's preferences (state of the negotiation and conditions), and Agent's and estimated partner's goals (actions), see Table 5.3. The Agent assumes that the partner's preferences are comparable to his, but values may differ. At the beginning of the interaction, the Agent may have no or weak assumptions about the partner's preferences. As the interaction proceeds the Agent builds up more knowledge about the partner's negotiation options. The Agent achieves this by taking the perspective of its partner and using its own knowledge to evaluate the partner's strategy, i.e. apply ToM skills. The Agent's memory holds three sets of preference values: the Agent's own preferences (zero ToM), the Agent's beliefs about the user's preferences (first-order ToM), and the Agent's beliefs about the user's beliefs about the Agent's preference values (second-order ToM).

When a negotiation move and its arguments are recognised, the information is passed to the CTA. The Agent constructs a retrieval instance and fills in as many slots as it can with the received details and the current context. Subsequently, the CTA updates its own

Information type	Explanation	Source
Strategy	The strategy associated with the instance	negotiationMove, modality
My-bid-value-me	The number of points the agent's bid is worth to the agent	
My-bid-value-opp	The number of points that the agent believes its bid is worth to the user	
Opp-bid-value-me	The number of points the user's bid is worth to the agent	
Opp-bid-greater	*true* if the user's bid is at least as much as the agent's current bid, *false* otherwise	Preference profile
Next-bid-value-me	The number of points that the next best option is worth	
	The next best option is defined as the option closest in value to the current one	
	(Not including those that are worth more than the current option.)	
Overall-value	The total value of all options that have been agreed upon so far.	
	This is a measure of how the negotiation is going.	History
	If it is negative, negotiation is likely to result in an unacceptable outcome.	
My-move	The move that the agent should take in this context.	Planned future

Table 5.3: Structure of an instance in the Cognitive Task Agent, adopted from [Stevens et al., 2016a].

representation of the negotiation state by retrieving an instance that has the highest *activation* value from CTA's declarative memory. An instance *i* that is used most recently and most frequently gets the highest activation value, or if no perfect match to a retrieval request is found a partial matching value is computed. Activation functions are derived from the equations presented in Section 3.5, also see [Bothell, 2004].

For example, suppose the CTA retrieves the following instance:

instance-a		
strategy	cooperative	the opponent's strategy is cooperative
my-bid-value-me	4	the agent's current offer is worth 4 points to him
opp-bid-value-me	1	the opponent's offer is worth 1 point to the agent
opp-bid-greater	true	the opponent's offer is equal or greater than agent's current bid
next-bid-value-me	2	the next best option for the agent is worth 2 points
opp-move	concede	opponent changed its offer to one that was less valuable to him
my-move	concede	the agent repays the opponent by also selecting a less valuable option

Two pieces of information from these instances will be extracted: the strategy of the user (cooperative) and an estimate of the user's preference for the options mentioned in the move. If there are other good options available, a cooperative negotiator will explore those options first before insisting on his current position, so from this behaviour the Agent infers that it is dealing with a cooperative negotiator with positive preferences on at least two issues. Now the Agent uses its own context to choose an appropriate response to the user. Depending on how the user has acted, and what the Agent knows (guesses) about the user's preferences, the Agent chooses to respond cooperatively, i.e. to concede.

5.4.3 Agents' multitasking behaviour

The designed Cognitive Task Agents can perform multiple tasks simultaneously. For example, the Debate Coach Agent can reason about the overall state of the debate task, and provides 'in- and about-action feedback on the debaters' performance concerning the presentation of arguments and debate success in achieving participants' goals. The Negotiation Agent can reason about the overall state of the negotiation task, and attempts to identify the best negotiation move for the next action. It computes: (1) the Agent's countermove, and (2) feedback sharing the Agent's beliefs about the user's preferences and the

Figure 5.6: Presentation Trainer interface mockup (right) and Freestyle Mode Interface (left). Adopted from Schneider et al. 2015.

user's negotiation strategy. The Agent may propose a strategically better alternative move that the user could have taken and explain 'why'. As the result, the system is able to play simultaneously or interchangeably the four roles specified in Section3.2: Observer, Negotiator, Mirror and Tutor.

In the Observer mode, both Agents monitor and keep track of all performed own and particiapants' actions and logs them. The created log files are used to evaluate the participants' performance and for system improvement (see next chapter).

As a Mirror, the Agents' monitoring and interpretation results are immediately displayed to the user. These displays include the transcripts of the Agents' and users' utterances (as recognised by the system), the Agents' perceived task related strategies and the recognised partners' behaviours and actions. In negotiation, the Agent's and partner's most recent offers and estimated partner's preferences are also flagged in the dynamically updated preference card (Figure 4.9). In debate, presentation behaviour is mirrored (visualised) to debaters. This has certain tutoring effects as well, since it activates/prompts learner's monitoring, reflection and regulating strategies, and triggers learner's corrective actions, also corrective Allo-Feedback acts on the Agent's processing failures, see Figure 5.6 for Presentation Trainer design and interfaces.

As a Negotiator[11], the Agent takes into account the recognised partner negotiation strategy, the Agent's preferences, and the estimation of those of the partner, and computes the most appropriate next negotiation move. This leads to relevant updates in the Semantic Context that give rise to goals to perform a certain dialogue act, e.g. tentative Agreement. Other contexts may also be updated in parallel and goals are created to perform, for example, turn-taking (Linguistic Context) and feedback (Cognitive Context) actions, see next section. The Dialogue Manager passes a list of dialogue acts for generation,

[11] Our system is not able to perform as an active debate partner yet. It is still a long way to a fully automatic and robust system that is able to understand debate arguments with high accuracy and to replace one of the debaters. This will be natural research continuation.

$< DA_1 = turnTake, DA_2 = positiveAutoFeedback, DA_3 = Task; Agreement >$, where DA_1 is decided to be generated implicitly, DA_2 - non-verbally by a smiling and nodding avatar and verbally by 'okay', and DA_3 is generated by the utterance 'I can live with it'.

As a Tutor, the Agents share their beliefs about the current negotiation or debate state. The Negotiation Agent can plan the negotiation ahead, e.g. may offer strategically better user negotiation moves leading to higher quality negotiation outcomes in terms of Pareto efficiency. After each action, the agent is also able to provide an explanation why decisions are made to perform certain actions. The Debate Coach Agent provides real-time feedback on presentational audible and visual performance. At the end of each negotiation session, summative feedback is generated in terms of the estimated Pareto optimality, degree of co-operativeness, and acceptance of negative outcomes. This type of feedback accumulates across multiple consecutive negotiation rounds. At the end of each debate session, summative feedback is generated in the form of all arguments discussed with indications of which of them are accepted and how the outcome deviates from the debaters' initial minimal goals.

The execution of these shared and varied tasks is expected to have positive effects both on user and system performance, enabling activation and improvement of metacognitive processes. Moreover, since these processes do not require additional resources (memory, processing and control), but are model-inherent belief creation and transfer processes and characteristics (instances slots), multiple tasks related to various roles can be executed by the DM in parallel without interference.

5.5 Dialogue Control Agents

Task actions account for less than half of all actions in our negotiation data, see Table 4.5. Other frequently occurring acts are concerned with Task Management, Discourse Structuring, Feedback and Social Obligations. Along with moving towards a final set of agreements, negotiators need to take care how to optimally structure and manage the negotiation and the interaction. In multi-issue bargaining, negotiators have a variety of task management strategies. They may discuss issues sequentially or bargain simultaneously about multiple options, making trade-offs across issues. They may withdraw and re-negotiate previously reached agreements. All these decisions require explicit communicative actions. The Task Management acts are recognised and generated by the system, and are modelled as part of the system's Semantic Context containing, along with the information about the speaker's beliefs about the negotiation domain, information concerning task progress and success. A Task Planner as part of the Task Manager (see Figure 5.1) takes care of updates and generation processes of this type.

Acts related to a negotiators' perception of the partner's physical presence and readiness to start, continue or terminate the interaction as well as participants' beliefs concerning the availability and properties of communicative and perceptual channels are modelled as part of the Perceptual Context. Dialogue behaviour addressing these aspects is important, in particular, these actions are considered for generation, since the system's multimodal behaviour related to Contact Management is embodied by a virtual character (full body avatar).

Processing level	Latest dialogue act		Previous dialogue act	Planned dialogue act
	Communicative Function	Negotiation Move		
Perception	unknown	unknown	any	Request Repeat and/or AutoNegative
Interpretation	unknown	offer(x)	any	Accept(offer(x)) or Reject(offer(x))
	unknown	offer(unknown)	any	Question(offer(?)) and/or AutoNegative
	any	unknown	any	Request Repeat or Rephrase and/or AutoNegative
	unknown	unknown	Accept(offer(x)) or Reject(offer(x))	Question(offer(?)) or Inform(offer(y))
	unknown	unknown	offer(x)	Question(offer(y))
	Question	unknown	Accept(offer(x)) or Reject(offer(x))	Inform(offer(y))

Table 5.4: Decision-making support for the system's feedback strategies concerning perception and interpretation of task-related actions, and expected dialogue continuation. Note: $x \neq y$.

The Contact Manager takes care of updates and the generation of these acts. Participant's beliefs concerning the interaction structure (i.e. history, present and future states) and beliefs concerning topic shifts are modeled as a part of the Linguistic Context; the Discourse Structuring module takes care of the updates and generation specific for the interaction management and monitoring.

5.5.1 Validity checking, repair and clarification strategies

For an interactive system it is important to know that its contributions are understood and accepted by the user, as well as to signal the system's processing of the same kind. Conversation is a **bilateral process** - that is, a joint activity, and speaking and listening are not autonomous processes - conversational partners monitor their own processing of the exchanged utterances as well as the processing done by the others, see Clark and Krych (2004) for discussion. Given the bilateral nature of conversation, interlocutors can construct and provide feedback on both their own processing (*auto-feedback*) as and on that by the other (*allo-feedback*).

Feedback is crucial for successful communication. Feedback can be provided at different **levels** of processing the communicative behaviour of interlocutors. Allwood et al. (1993) and Clark (1996) notice that interlocutors need to establish *contact* and gain or pay *attention* to each others behaviour, in order be involved in conversation. A speaker's behaviour needs to be *perceived* (i.e. heard, seen) or *identified* (Clark, 1996). Perceived behaviour should be *interpreted*, i.e. interlocutors should be able to extract the meaning of each other's behaviour. The constructed interpretation needs to be *evaluated* against one's information state: if it is consistent with the current information state it can be incorporated into that state; if it is inconsistent, this can be reported as negative feedback. The incorporation of new information, and the performance of other mental and physical actions in response to communicative behaviour is called the *execution* or *application* (Bunt, 2000).

Processing level	Latest dialogue act		Previous dialogue act	Validity	Planned dialogue act
	Communicative Function	Negotiation Move			
Evaluation	Inform	terminate	any	valid	stop negotiation
	Accept	offer(x)	final Offer(x)	valid	Inform(deal(x))
	any other than Accept	offer(x)	final Offer(x)	invalid	Auto/AlloFeedback: Question(?offer(x))
	Accept	deal(x))	Inform(deal(x))	valid	DiscourseStructuring: TopicShift TaskManagement:Suggest(next_issue) DiscourseStructuring: Closing
	Reject or Accept	offer(x)	Suggest(offer(x))	valid	Inform(offer(y)); Question(offer(y))
	Reject or Accept	offer(x)	Inform(offer(x))	valid	Inform(offer(y)); Question(offer(y))
	Question	offer(x)	Inform(deal(x))	valid	Inform(deal(x))
	Reject	offer(x)	Reject(offer(x))	valid	Reject(offer(x))
	final Offer or Accept	offer(x)	Accept(offer(x))	valid	Accept(offer(x))
	Inform Suggest Offer	offer(x)	Accept(offer(y))	valid if x= ¬y	Accept(offer(?x)) or Reject(offer(x))
	Inform	offer(y)	Inform(deal(x))	invalid	interpret as Accept(deal(x)) if x=y otherwise Reject(offer(x))
	Accept	offer(x)	any other than Suggest or Inform(offer(x))	invalid	interpret as Inform(offer(x)) or Suggest(offer(x)) if x=y otherwise generate Auto- or AlloNegative
	Reject	offer(x)	any other than Suggest or Inform(offer(x))	invalid	interpret as Inform(offer(y)) or Suggest(offer(y)) if x¬y otherwise generate Auto- or AlloNegative
	Accept	offer(x)	Inform(terminate)	invalid	interpret as Accept(terminate) and generate Auto- or AlloNegative
	Reject	offer(x)	Inform(terminate)	invalid	interpret as Accept(terminate) and/or generate Auto- or AlloNegative
	Inform	deal(x)	Inform(terminate)	invalid	Question(offer(?)) and/or Auto- or AlloNegative
	Inform Suggest Offer	offer(x)	Accept(offer(y))	invalid if x=y	Question(offer(?x)); Accept(offer(y)) and/or Auto- or AlloNegative
	Inform Suggest Offer	offer(x)	Reject(offer(y))	invalid if x=y	Question(offer(?x)); Reject(offer(y))and/or Auto- or AlloNegative
	Inform	deal(x)	Reject(offer(x))	invalid	Reject(offer(x)); Question(offer(?x)) and/or Auto- or AlloNegative

Table 5.5: Decision-making support for the system's recovery and clarification strategies concerning evaluation of task-related actions, and expected dialogue continuation. In this table, *valid* stands for the state that can be recovered from the available information, otherwise *invalid* - state that cannot be automatically recovered and requires activation of the clarification strategy. Note: $x \neq y$.

Processing level	Latest dialogue act		Previous dialogue act	Preferences	Validity	Planned dialogue act
	Communicative Function	Negotiation Move				
Execution	Inform Suggest Offer	offer(x)	any	negative	valid	Reject(offer(x)) and/or Inform(offer(y)) and/or AutoNegative
	Inform Suggest Offer	offer(x)	any	positive	valid	Accept(offer(x)) and/or Inform(offer(y)) and/or AutoNegative
	Inform Suggest Offer	offer(x)	any	neutral	valid	Accept(offer(x)) and/or Inform(offer(y))
	Inform	deal(x)	no Accept(offer(x))	any	invalid	Reject(deal(x)); Question(offer(?x)) and/or AutoNegative
	Inform	deal(x)	Reject(offer(x))	any	invalid	Reject(deal(x)); Question(offer(?x)) and/or AutoNegative
	inform	terminate	no final offer(x)	any	invalid	Question(offer(?x)) and/or AutoNegative

Table 5.6: Decision-making support for the system's feedback strategies concerning execution of task-related actions. In this table, *valid* stands for the state that can be recovered from the available information, otherwise *invalid* - state that cannot be automatically recovered and requires activation of the clarification strategy. Note: $x \neq y$

A speaker may provide feedback (***feedback giving***) or elicit feedback (***feedback eliciting***).

As for positive feedback acts, explicitly signalled acceptances are generated, either verbally or non-verbally. We also consider generation of multimodal expressions of implied and entailed positive feedback (see [Bunt, 2007, Bunt, 2012]) for strategic reasons, e.g. to provide more certainty due to potentially erroneous automatic speech recognition output.

Detected difficulties and inconsistencies in recognition, interpretation, evaluation and execution need to be resolved immediately if these problems are serious enough to impede further task performance; such problems are reported accordingly. Problems due to deficient recognition and interpretation are frequent in spoken human-computer dialogue systems, but rarely observed in the collected human-human dialogue data. Good news however is that humans generally exhibit certain re-occurring behavioural patterns when their processing fails. For our scenario and dialogue setting we incorporated observations and analyses of other available dialogue resources such as the human-human AMI and HCRC MapTask corpora (Carletta, 2006; Anderson et al., 1991), and human-human and human-computer DBox quiz game data (Petukhova et al., 2014; 2015).

ID	Utterance *(wording)*	DM input: DA ID	DM input: D;CF [dependence]	DM input: SC=NM($I;V$)/ modality	DM Information State: CTA state /decision	DM Information State: DM update[12]	DM output: DA for generation
A_1	what do you want for scope	da1		p1=offer(1;?v)	neutral/elicit	$Wants(A, Know(A, p1))$	task;setQuestion
C_1	i think it would be fine					$Bel(C, Wants(A, Know(A, p1)))$	
	if we stop smoking			p2=offer(1;b)/		$Bel(C, \Box p2)$	
	in public transportation	da2	task;answer[da1]	prefer	cooperative	$Wants(C, Know(A, \Box p2))$	
$A_{2.1}$	okay					$Bel(A, Wants(C, Know(A, \Box p2))$	
$A_{2.2}$	i would go for that point				cooperative/	$Bel(A, \neg\Box p2))$	
		da3		p2=offer(1;b)	agree(A,1b)	$Wants(A, Know(C, \Diamond p2))$	task;agreement
$A_{2.3}$	i prefer that we do not				cooperative/		
	increase taxes	da4		p3=offer(2;a)	offer(A,2a)	$Wants(A, Know(C, \Box p3))$	task;inform
$C_{2.1}$	this is not possible	da5	task;	p3=offer(2;a)	non-cooperative/	$Bel(C, Wants(A, Know(C, \Box p3)))$	
	for us		disagreement[da4]	unable	reject(C,2a)	$Wants(C, Know(A, \neg\Diamond p3))$	
$C_{2.2}$	i would like 5% tax						
	increase at least	da6	task;inform	-	-	-	
A_3	sorry i was distracted						apology &
	could you repeat	da7				$Bel(A, \neg Interpreted(A, C_{2.2}))$	autoNegative
C_3	sure	da8	autoPositive[da7]			$Bel(C, Interpreted(C, A_3))$	
	I prefer 5% tax increase	da9	task;inform	p4=offer(2;b)	non-cooperative/	$Bel(C, Wants(A, Know(C, \neg\Diamond p3))$	
				prefer	offer(C,2b)	$Wants(C, Know(A, \Box p4))$	
$A_{4.1}$	okay	da10	autoPositive[da9]	p4=offer(2;b)		$Bel(A, Interpreted(A, C_3))$	autoPositive
$A_{4.3}$	i will give you 5% if you			p4=offer(2;b)		$Bel(A, Wants(C, Know(A, \Box p4))$	
						$Bel(A, \Diamond p4)$	
	agree to ban all tobacco			p5=offer(4;b)	neutral/	$Bel(A, \Box p5)$	
	vending machines	da11		p6=offer(	exchange(A,	$Bel(A, \Diamond p6)$	task;offer
				(2;b) $\rightarrow$ (4;b))	(2b $\wedge$ 4b))	$Wants(A, Know(C, \Diamond p6))$	
C_4	i think i can live with that	da12	task;	p6=offer(	cooperative/agree(C,	$Bel(C, Wants(A, Know(C, \Diamond p6))$	
			agreement[da11]	(2;b) $\rightarrow$ (4;b))	(2b $\rightarrow$ 4b))	$Bel(C, \Diamond p6)$	

Table 5.7: Example of a negotiation dialogue with processing and generation by the Dialogue Manager. (A = agent (Business Representative); C = human negotiator (City Councilor); DA = dialogue act; D = dimension; CF = communicative function; SC= semantic content; NM = negotiation move; I = issue; V=value; Bel = believes; $\Diamond$ = possible; $\Box$ = preferable)

Observations from human-human and human-computer dialogues resulted in the definition of feedback strategies at the level of perception (recognition) and interpretation mostly comprising corrections and requests to repeat or rephrase (Table 5.4), at the level of evaluation reporting inconsistencies/(in)validity due to certain logical constraints, given the grounded negotiation history (Table 5.5), and at the level of execution reporting inability to accept an offer or to reach an agreement due to the negotiator's preference profile (Table 5.6). Certain system processing flaws can be recovered from the information available to the system, some problems are too severe to continue the dialogue successfully and trigger feedback acts (clarification requests). In total, about 30 clarification and recovery strategies have been defined and evaluated (see also next Chapter).

Information concerning successes and failures in the processing of a partners' dialogue contributions are modelled as part of the Cognitive Context (see Figure 5.3). Table 5.7 provides an example of a dialogue between an agent *A* playing the role of the Business Representative and a human negotiator *C* in the role of the City Councilor. The Negotiation Agent starts neutrally. *A* elicits an offer from *C* on the first issue and does this in the form of a Set Question. The understanding that a certain dialogue act is performed leads to corresponding context model updates. If the partner reacts to the agent's elicitation by sharing his preferences in C_1, he is evaluated by the agent as being cooperative. The agent's preferences are not identical but not fully conflicting either: it is possible for the agent to agree with the opponent's preferences accepting his offer in $A_{2.2}$, where *A* believes that the offer made in C_1 is not the most preferred one but still acceptable/possible for *A*.[12] The NA stays in the cooperative mode. If the negotiator's preferences differ from the options proposed by the partner, he may refuse to accept the partner offer as in $C_{2.1}$ and may offer another value which is more preferable for him, i.e. perform a counter-offer move ($C_{2.2}$ repeated in C_3 after the agent signaled that his processing was unsuccessful. The NA interprets the partner's strategy as being non-cooperative and switches his strategy to neutral, proposing to exchange offers (in $A_{4.2}$) that still aim at the better deal for himself. If this will again be rejected, the agent will apply the non-cooperative strategy and insist on his previous proposal expressed in $A_{2.2}$, otherwise he will either elicit an offer for the next issue or propose an offer himself.

Dialogue control acts present an important part for any interaction. In a shared cultural and linguistic context, choices concerning the frequency of such actions and the variety of expressions are rather limited. Conventional forms are mostly used to greet each other, to apologize, to manage the turns and the use of time, to deal with speaking errors, and to provide or elicit feedback. Models of dialogue control behaviour once designed can therefore be applied in a wide range of communicative situations. The use of task-related dialogue acts, by contrast, is more application-specific. The separation between task-related and dialogue control actions is therefore not only a cost-effective solution, but also allows designing flexible architectures and combinations of different modelling approaches and techniques, resulting in more robust and rich system behaviour.

[12] We provide here a simplified representation of the participants' information states as tracked and updated by the DM. The full specification of participants' information states and their updates can be found in Table 5.9.

5.6 Dialogue Manager state update and belief transfer

To be successful in communication, the participants have to coordinate their activities on many levels. In the speaker role, a participant produces utterances with the aim to be understood by others. In dialogue act theory, understanding that a certain dialogue act is performed means *creating* the belief that the preconditions hold which are characteristic for that dialogue act. This not only holds for 'private' individual beliefs of a participant. All participants involved in the interaction work collaboratively on coordination of the beliefs and assumptions of the participants and this coordinating activity is central to any communication. A set of propositions that the dialogue participants *mutually* believe is called their *common ground*, and the process of establishing and updating the common ground is called *grounding*. The speaker expects under 'normal input-output' conditions [Searle, 1969] that what he is saying is perceived and understood as intended. These expectations may be *strengthened* when there is positive evidence from the audience, and if negative feedback occurs the expectations are *canceled*. Evidence for belief strengthening and cancellation often take the form of explicit or implicit positive feedback. Not all propositions are addressed immediately, and grounding may be postponed/delayed. A participant does not only expect to be understood, but works towards the goal that the addressees incorporate his beliefs as beliefs of their own (belief *adoption*), e.g. a debater wants to convince his audience of the rightness of his position.

Processing of the incoming utterance several stages can be distinguished: awareness, recording, buffering, acceptance, and adoption. Each processing stage corresponds to the application of a number of general mechanisms for updating the addressee's context that can be defined as follows:

- **Creation**: an interlocutor introduces a belief as the effect of assigning an interpretation to what has been said by another interlocutor. Creation has two stages: (1) addition of precondition to the pending (or temporary) context; and (2) acceptance of beliefs and addition of accepted elements to the main context;
- **Adoption**: an interlocutor incorporates beliefs of an other interlocutor as beliefs of his own;
- **Cancellation**: a belief or goal is cancelled because it does not apply any more, or a goal has been achieved or has been understood to be unachievable;
- **Strengthening**: an expectation, or 'weak belief' becomes a firm belief because sufficient supporting evidence for the belief becomes available.

5.6.1 Belief transfer in debate

In parliamentary debates, where political confrontations and ideological convictions often play a significant role, the goals of a debater depend on the type of debate. In legislation debates the main goal is to gain the majority of supporters in terms of votes. A lot of preparatory work is done before the actual debate takes place, in committees and lobbies.

To achieve their main goal parliamentarians may be ready to compromise on some points and negotiate on others. A governing party with a majority in the parliament has a bigger chance to get their beliefs adopted by the majority, therefore has stronger initial expectations. Parliamentarians also have certain knowledge about their opponents and their seconders, which should be modelled in the initial dialogue context together with knowledge about common and individual goals, and should be taken into consideration when computing the strength of expectations concerning the outcome of a debate. In HCI research it is common to incorporate user models where all available information about dialogue participants is specified [Fischer, 2001]. This type of information is typically useful to design adaptive human-computer systems and can be profitably used when modelling interactive behaviour in dialogue, in particular related to grounding.

In many debate situations, no strong political division is obvious a priori, and it is reasonable to assume that each debater expects that many of his partners will adopt his beliefs. At least, this is what he strives for, otherwise it would make little sense to participate in such a debate. With this goal in mind, a participant does his best to be convincing and persuasive, presenting his claims and evidence as convincingly as possible. Debater D_1 thinks that to forbid smoking in all public places including open air areas is inappropriate. The debaters D_2, D_3 understand this proposition and make it part of their common ground. Debater D_2 disagrees with D_1's position presented in the argument 1.2 and attacks it in 2.1. Thus, D_1 expected adoption effect is cancelled. Debater D_3, by contrast, supports D_1's argument 1.2 and suggests to forbid smoking inside all public places but not in open air areas. As a result, the D_1 expected adoption effect holds and that of D_2 is cancelled. The participants' beliefs are updated as follows, where *Bel* stands for believes, *MBel* mutually believed and *WBel* for weakly believes:

(13) D1_1.2:*We should forbid smoking in public places, but to forbid smoking in open air areas goes too far.*
preconditions:
$Bel(D_1, arg1.2)$; $Want(D_1, Bel(\{D2, D3\}, arg1.2))$
expected understanding:
$Bel(D_1, MBel(\{D_1, D_2, D_3\}, WBel(D_1, Bel(\{D2, D3\}, Bel(D_1, arg1.2)))))$;
$Bel(D_1, MBel(\{D_1, D_2, D_3\}, WBel(D_1, Bel(\{D2, D3\}, Want(D_1, Bel(\{D2, D3\}, arg1.2))))))$
expected adoption:
$Bel(D_1, MBel(\{D_1, D_2, D_3\}, WBel(D_1, Bel(\{D2, D3\}, arg1.2))))$

D2_2.1: *We should stay firm and not make any exceptions, if smoking is forbidden, then in all public places.*
understanding:
$MBel(\{D_1, D_2\}, Bel(D_1, arg1.2))$; $MBel(\{D_1, D_2\}, Want(D_1, Bel(D_2, arg1.2)))$
cancelled adoption:
~~$Bel(D_1, MBel(\{D_1, D_2\}, WBel(D_1, Bel(D_2, arg1.2))))$~~
preconditions:
$Bel(D_2, \neg arg1.2)$; $Want(D_2, Bel(\{D_1, D_3\}, \neg arg1.2))$
expected understanding:
$Bel(D_2, MBel(\{D_1, D_2, D_3\}, WBel(D_2, Bel(\{D_1, D_3\}, Bel(D_2, \neg arg1.2)))))$;

$Bel(D_2, MBel(\{D_1, D_2, D_3\}, WBel(D_2, Bel(\{D_1, D_3\}, Want(D_2, Bel(\{D_1, D_3\}, \neg arg1.2))))))$
expected adoption:
$Bel(D_2, MBel(\{D_1, D_2, D_3\}, WBel(D_2, Bel(\{D_1, D_3\}, \neg arg1.2))))$

D3_3.1: *I think allow smoking in open air places is perfectly sensible.*
understanding:
$MBel(\{D_1, D_3\}, Bel(D_1, arg1.2))$; $MBel(\{D_1, D_3\}, Want(D_1, Bel(D_3, arg1.2)))$;
$MBel(\{D_3, D_2\}, Bel(D_2, \neg arg1.2))$; $MBel(\{D_3, D_2\}, Want(D_2, Bel(D_3, \neg arg1.2)))$
adoption:
$Bel(D_1, MBel(\{D_1, D_3\}, arg1.2))$
cancelled adoption:
~~$Bel(D_2, MBel(\{D_2, D_3\}, WBel(D_2, Bel(D_3, \neg arg1.2))$~~
preconditions:
$Bel(D_3, arg1.2)$; $Want(D_3, Bel(\{D_1, D_2\}, arg1.2))$
expected understanding:
$Bel(D_3, MBel(\{D_1, D_2, D_3\}, WBel(D_3, Bel(\{D_1, D_2\}, Bel(D_3, arg1.2)))))$;
$Bel(D_3, MBel(\{D_1, D_2, D_3\}, WBel(D_3, Bel(\{D_1, D_2\}, Want(D_3, Bel(\{D_1, D_2\}, arg1.2))))))$
expected adoption:
$Bel(D_3, MBel(\{D_1, D_2, D_3\}, WBel(D_3, Bel(\{D_1, D_2\}, arg1.2))))$

The Dialogue Manager keeps track of all created and adopted beliefs on the part of each debater as the debate proceeds. We used the conclusions identified in the presented ADUs to update the information states of participants and that of the system. To give an example, the debate outcome looks as illustrated in (14):

(14) Arg_1: Not all public places should be affected, allow smoking in open air areas [*Support 1.1, 3.1, 4.1, 6.1/Attack 2.1, 5.1*]
Arg_2: Tobacco price already high, no tax increase necessary [*Support 1.2, 5.2/ Attack 2.2, 3.2, 4.2, 6.2*]
Arg_3: Tobacco should be sold in supermarkets and specialised licensed tobacco shops [*Support 1.3, 3.3, 4.3, 6.3/Attack 2.3, 5.3*]
Arg_4: No police control but municipal and administrative control, no penalties but warnings for the 1st time disobedience [*Support 1.4, 3.4, 4.4, 5.4, 6.4/ Attack 2.4*]
Arg_5: An anti-smoking campaign should involve all mass media channels TV [*Support 1.5, 2.5, 3.5, 4.5, 5.5, 6.5*]

This leads to the system creating and adopting beliefs concerning these arguments. For example, with regard to the argument $Arg_{4.1}$ in (14) the following system beliefs are created: $Bel(S, MBel(\{S, D_1, D_3\}, Bel(\{D_1, D_3\}, Arg_{4.1})))$, $Bel(S, MBel(\{S, D_1, D_3\}, Want(\{D_1, D_3\}, Bel(S, Arg_{4.1}))))$, where S stands for System. In the final state, the system may predict that the belief $Bel(S, MBel(\{S, D_1, D_3\}, Arg_{4.1}))$ will be adopted and will constitute a part of the debate outcome. The final predicted system state is compared with the actual outcome summarised by human tutors and debaters (see Section 5.7).

For the system in the tutoring role training the trainee presentational skills, the systems is expected to understand the trainees' behaviour and judge its appropriateness in the debate

or negotiation situation based on the criteria defined in Table 5.1. When inappropriate behaviour is detected, the system reacts by either informing the addressee of his infelicitous use of voice and body, or propose how this behaviour can be corrected. At the same time the system does not want to take the turn over, but rather communicate its messages in a non-intrusive manner. Thus, system responses are generated in visual form using colours (red meaning something wrong happened, green - participant's performance is according to expectations) and pictures depicting correct body position, plus verbal written message with the actual system verbal response.

The context model is updated as shown in Table 5.8.[13] The system believes to have interpreted participant P1's dialogue contribution having certain semantic content (representation of verbal component, speakingVolume:HIGH and gesture:ARMSCROSSED which P1 believes are correct) and communicative function (Inform). Using the knowledge available to the system, e.g. in a database with prosodic properties and visible body movements that are inappropriate in a debate situation (see also Table 5.1 for illustration), the system's task is to inform the addressee about his presentational failures. Thus, the system has a choice to generate either a Task Inform act with content $\neg appropriate(volume(fs_1) = high)$ and $\neg appropriate(gesture(fs_1) = ARMSCROSSED)$, or a Task Correction act as illustrated in Table 5.8. The system expects that its dialogue acts are successfully interpreted (s2, s3) and adopted (s07, s08) by the participant P1, and when continuing the dialogue he will lower his voice volume and uncross his arms (adoption in u08 and u09 leading to dialogue acts da_4 and da_5 expressed in one multifunctional functional segments fs_4). Formally defined update operators for these dialogue acts can be found in Bunt (2014).

If, by contrast, the addressee does not recognise the system's dialogue acts or is not able to perform corrected actions, this will lead to *cancellation* of expected adoption effects. A belief or goal is further cancelled for this participant because it does not apply any more. Cancellation of a goal will also occur when the goal has been achieved or has been understood to be unachievable. Weak beliefs can be *strengthened* later as supporting evidence becomes available (see also Bunt et al., 2007).

[13]NOTE: For the sake of simplicity we do not spell out the updates on all debate participants' states (the opponent's behaviour interpretation is not required for this task). The example serves to illustrate the underlying general principles.

Table 5.8: Example of context update. (LC = Linguistic Context; CC = Cognitive Context; SC = semantic context; prec = preconditions; du = dialogue utterance; da = dialogue act; fs = functional segment; D = dimension; CF = communicative function; exp.und = expected understanding; und = understanding; exp.ad = expected adoption; ad = adoption; Bel = believes; MBel = mutually believed; WBel = weakly believes)

Context	num	source	S's context	num	source	P_1's context
LC				u001	*prec*	$Bel(P_1, Next\_Speaker(P_1))$
LC	s1		$Bel(S, Current\_Speaker(P_1))$	u1		$Bel(P_1, Current\_Speaker(P_1))$
	$fs_1 : du1$	*latest*	$\langle verbatim \rangle$	$fs_1 : du1$	*latest*	$\langle verbatim \rangle$
			$volume(fs_1) = high$			$volume(fs_1) = high$
			$gesture(fs_1) = armscrossed$			$gesture(fs_1) = armscrossed$
	$fs_1 : da_1$	*D;CF*	Task; Inform	$fs_1 : da_1$	*D;CF*	Task; Inform
		sem_content	$\langle p \rangle$		*sem_content*	$\langle p \rangle$
			Speaker:P_1; Addressee: P_2			Speaker:P_1;Addressee: P_2
SC	s01a	exp.und:$fs_1 : da_1$	$Bel(S, MBel(\{S, P_1\}, WBel(P_1,$	u01a	exp.und:$fs_1 : da_1$	$Bel(P_1, MBel(\{S, P_1\}, WBel(P_1,$
			$Bel(S, Bel(P_1, appropriate($			$Bel(S, Bel(P_1, appropriate($
			$volume(fs_1) = high)))$			$volume(fs_1) = high)))$
	s01b	exp.und:$fs_1 : da_1$	$Bel(S, MBel(\{S, P_1\}, WBel(P_1,$	u01b	exp.und:$fs_1 : da_1$	$Bel(P_1, MBel(\{S, P_1\}, WBel(P_1,$
			$Bel(S, Bel(P_1, appropriate($			$Bel(S, Bel(P_1, appropriate($
			$gesture(fs_1) = armscrossed)))$			$gesture(fs_1) = armscrossed)))$
SC	s2a	*prec*	$Bel(S, \neg appropriate(volume(fs_1) = high))$			
			$Bel(S, appropriate(volume(fs_1) = medium))$			
	s3a		$Want(S, Bel(P_1, appropriate($			
			$volume(fs_1) = medium)))$			
	s2a	*prec*	$Bel(S, \neg appropriate(gesture(fs_1) = armscrossed))$			
			$Bel(S, appropriate(gesture(fs_1) = armsUncrossed))$			
	s3b		$Want(S, Bel(P_1, appropriate($			
			$gesture(fs_1) = armsUncrossed)))$			
LC	da_2	*plan:s03a*	Task; Correct			
		sem_content	$appropriate(volume(fs_1) = medium)$			
	da_3	*plan:s03b*	Task; Correct			
		sem_content	$appropriate(gesture(fs_1) = armsUncrossed)$			
LC	s04	*prec*	$Bel(S, Current\_Speaker(P_1))$			
			$Want(S, Next\_Speaker(P_1))$			

Context	num	source	S's context	num	source	P_1's context
LC	$fs_2 : du2$	*latest*	$\langle VOLUME\_MEDIUM \rangle$			
	$fs_2 : da_2$	*D;CF*	Task; Correct			
	$fs_3 : du3$	*latest*	$\langle UNCROSS\_ARMS \rangle$			
	$fs_3 : da_3$	*D;CF*	Task;Correct			
CC	s2	exp.und:da_2	$Bel(S, MBel(\{S, P_1\}, WBel(P_1, Interpreted(P_1, du_2)))$	u2	exp.und:da_2	$Bel(P_1, MBel(\{S, P_1\}), WBel(P_1, Interpreted(P_1, du2)))$
	s3	exp.und:da_3	$Bel(S, MBel(\{S, P_1\}, WBel(P_1, Interpreted(P_1, du_3)))$	u3	exp.und:da_3	$Bel(P_1, MBel(\{S, P_1\}), WBel(P_1, Interpreted(P_1, du3)))$
SC	s05	exp.und:da_2	$Bel(S, MBel(\{S, P_1\}, WBel(P_1, Want(S, Bel(P_1, appropriate($ $volume(fs_1) = medium))))))$	u02	exp.und:da_2	$Bel(P_1, MBel(\{S, P_1\}), WBel(P_1, Want(S, Bel(P_1, appropriate($ $volume(fs_1) = medium))))))$
	s06	exp.und:da_3	$Bel(S, MBel(\{S, P_1\}, WBel(P_1, Want(S, Bel(P_1, appropriate($ $gesture(fs_1) = armsUncrossed))))))$	u03	exp.und:da_3	$Bel(P_1, MBel(\{S, P_1\}), WBel(P_1, Want(S, Bel(P_1, appropriate($ $gesture(fs_1) = armsUncrossed))))))$
	s07	exp.ad:da_2	$Bel(S, MBel(\{S, P_1\}, WBel(P_1, Bel(P_1, appropriate($ $volume(fs_1) = medium)))))$	u04	exp.ad:da_2	$Bel(P_1, MBel(\{S, P_1\}), WBel(P_1, Bel(P_1, appropriate($ $volume(fs_1) = medium)))))$
	s08	exp.ad:da_3	$Bel(S, MBel(\{S, P_1\}, WBel(P_1, Bel(P_1, appropriate($ $gesture(fs_1) = armsUncrossed)))))$	u05	exp.ad:da_3	$Bel(P_1, MBel(\{S, P_1\}), WBel(P_1, Bel(P_1, appropriate($ $gesture(fs_1) = armsUncrossed))))$
SC				u06	und:da_2	$Bel(P_1, Want(S, Bel(P_1, appropriate($ $volume(fs_1) = medium))))$
				u07	und:da_3	$Bel(P_1, Want(S, Bel(P_1, appropriate($ $gesture(fs_1) = armsUncrossed)))$
				u08	ad:da_2	$Bel(P_1, appropriate($ $volume(fs_1) = medium))$
				u09	ad:da_3	$Bel(P_1, appropriate($ $gesture(fs_1) = armsUncrossed))$
LC				da_4	*plan:u08*	Task; Inform
					sem_content	$appropriate(volume(fs_1) = medium)$
				da_5	*plan:u09*	Task; Inform
					sem_content	$appropriate(gesture(fs_1) = armsUncrossed)$
LC				u002	*prec*	$Bel(P_1, Next\_Speaker(P_1))$
LC				$fs_4 : du4$	*latest*	$\langle verbatim \rangle$ $volume(fs_4) = medium$ $gesture(fs_4) = armsUncrossed$

5.6.2 Belief transfer in negotiations

In negotiations, negotiators aim at the understanding by others as well. Applying the ISU machinery and procedures, the same as for debates, to incorporate beliefs and expectations shared between speaker and hearers, we computed *expected understanding effects* modelled as *weak* beliefs. When evidence about successful understanding arrives weak beliefs are *strengthened*, otherwise they may be *cancelled*.

Negotiators also expect that their opponent will accept at least some of their offers (*expected adoption effects*). The strength of such expectations depends on their knowledge about their opponents, on their goals, and on the knowledge concerning the opponent's negotiation strategy. When the negotiator states identical preferences, agrees with the opponent's preferences, or accepts his suggestions and offers, he adopts the opponent's beliefs as beliefs of his own. For example:

(15) Council(human): *What do you think if we do not allow smoking in public transportation at least?*
Business(agent): *Well, I think we can live with that*

Council (C) produces a $\langle Task; suggest \rangle$ dialogue act with the semantic content $p2$. Weak mutual beliefs concerning expected understanding and adoption effects are created, the dialogue context model is updated with $s01a - s02c$ and $u01a - u02c$ updates as shown in Table 5.9. Business representative A understands C's da_1 as a suggestion and accepts it following the cooperative negotiation strategy. A's understanding means that A believes that C wants A to consider to do $p2$ because C believes that $p2$ would be interesting for A, and A is able to do $p2$. In A's preference profile, $p2$ is a possible offer. This enables A to accept C's suggestion, see precondition in $s3$. A acting as a cooperative agent is considering to offer the discussed value and commits to perform this action. Thus, beliefs about expected and actual understanding and adoption together with the negotiator's preferences give rise to the generation of one or more relevant dialogue acts. Similarly, additional updates are performed in other contexts. For instance, the Linguistic Context (LC) is updated with respect to beliefs concerning the speaker role management, and in the Cognitive Context (CC) concerning processing successes and failures. This triggers the generation of dialogue acts in multiple dimensions, e.g. here in the Turn Management and Feedback dimensions, respectively.

Table 5.9: Example of context update for cooperative negotiation behaviour. (LC = Linguistic Context; CC = Cognitive Context; SC = semantic context; prec = preconditions; da = dialogue act; fs = functional segment; D = dimension; CF = communicative function; exp.und = expected understanding; und = understanding; exp.ad = expected adoption; ad = adoption; Bel = believes; MBel = mutually believed; WBel = weakly believes)

Context	num	source	Agent (A)context	num	source	Council (C) context
LC				u001	*prec*	$Bel(C, Next\_Speaker(C))$
LC	s1		$Bel(A, Current\_Speaker(C))$	u1		$Bel(C, Current\_Speaker(C))$
	fs_1	*latest*	$\langle verbatim \rangle$	fs_1	*latest*	$\langle verbatim \rangle$
	da_1	*D;CF*	Task; Suggest	da_1	*D;CF*	Task; Suggest
		sem_content	$p2 = offer(ISSUE = 1; VALUE = 1b)$		*sem_content*	$p2 = \langle offer(ISSUE = 1; VALUE = 1b)$
			Speaker:*C*; Addressee: *A*			Speaker:*C*;Addressee: *A*
CC	s2	exp.und:da_1	$Bel(A, MBel(\{A,C\}, WBel(C,$ $Interpreted(A, du_1))))$	u2	exp.und:da_1	$Bel(A, MBel(\{A,C\}, WBel(C,$ $Interpreted(A, du_1))))$
SC	s01a	exp.und:da_1	$Bel(A, MBel(\{A,C\}, WBel(C,$ $Bel(A, Bel(C, Interest(A, p2))))))$	u01a	exp.und:da_1	$Bel(C, MBel(\{A,C\}, WBel(C,$ $Bel(A, Bel(C, Interest(A, p2))))))$
	s01b	exp.und:da_1	$Bel(A, MBel(\{A,C\}, WBel(C,$ $Bel(A, Assume(C, CanDo(A, p2))))))$	u01b	exp.und:da_1	$Bel(C, MBel(\{A,C\}, WBel(C,$ $Bel(A, Assume(C, CanDo(A, p2))))))$
	s01c	exp.und:da_1	$Bel(A, MBel(\{A,C\}, WBel(C,$ $Bel(A, Want(C, ConsidDo(A, p2))))))$	u01c	exp.und:da_1	$Bel(C, MBel(\{A,C\}, WBel(C,$ $Bel(A, Want(C, ConsidDo(A, p2))))))$
	s02a	exp.ad: da_1	$Bel(A, MBel(\{A,C\}, WBel(C,$ $Bel(A, Interest(A, p2))))))$	u02a	exp.ad:da_1	$Bel(C, MBel(\{A,C\}, WBel(C,$ $Bel(A, Interest(A, p2))))))$
	s02b	exp.ad: da_1	$Bel(A, MBel(\{A,C\}, WBel(C,$ $Bel(A, CanDo(A, p2))))))$	u02b	exp.ad:da_1	$Bel(C, MBel(\{A,C\}, WBel(C,$ $Bel(A, CanDo(A, p2))))))$
	s02c	exp.ad: da_1	$Bel(A, MBel(\{A,C\}, WBel(C,$ $Bel(A, ConsidDo(A, p2))))))$	u02c	exp.ad:da_1	$Bel(C, MBel(\{A,C\}, WBel(C,$ $Bel(A, ConsidDo(A, p2))))))$
SC	s03a	und:da_1	$Bel(A, Bel(C, Interest(A, p2)))$			
	s03b		$Bel(A, Assume(C, CanDo(A, p2)))$			
	s03c		$Bel(A, Want(C, ConsidDo(A, p2)))$			
SC	s3	*prec*	$Bel(A, \Diamond p2)$			
	s04a	ad:da_1	$Bel(A, Interest(A, p2))$			
	s04b		$ConsidDo(A, p2)$			
	s04c		$Bel(A, CanDo(A, p2))$			
SC	s4	*prec*	$CommitDo(A, p2)$			
LC	da_2	*plan:s4*	Task; AcceptSuggest			
		sem_content	$p2 = offer(ISSUE = 1; VALUE = 1b)$			
LC	s001	*prec*	$Bel(A, Next\_Speaker(A))$			
LC	s5		$Bel(A, Current\_Speaker(A))$	u2		$Bel(C, Current\_Speaker(A))$
	fs_2	*latest*	$\langle verbatim \rangle$	$fs_2 : du2$	*latest*	$\langle verbatim \rangle$
	da_2	*D;CF*	Task;AcceptSuggest	$fs_2 : da_2$	*D;CF*	Task;AcceptSuggest
		sem_content	$p2 = offer(ISSUE = 1; VALUE = 1b)$		*sem_content*	$p2 = offer(ISSUE = 1; VALUE = 1b)$
			antecedent: da_1			antecedent: da_1
			Speaker:*A*; Addressee: *C*			Speaker:*A*;Addressee: *C*

The example in (16) shows non-cooperative negotiation behaviour. It may be noted that negotiation partners always cooperate at a linguistic level, as they try to understand each other's contributions and respond to perceived intentions.[14] A rational agent may show non-cooperative behavior at the level of perlocutionary actions (see [Attardo, 1997]), when cancelling of expected adoption beliefs occurs.

(16) Council(human): *What do you think if we do not allow smoking in public transport at least?*
Business(agent): *It's not possible for me*

The dialogue context model is updated in this case as follows. A understanding C means that A believes that C wants A to consider to do $p2$ because C believes that $p2$ would be interesting for A and A is able to do $p2$. According to A's preference profile, $p2$ is not a possible offer, resulting in the precondition in $s3$ as $Bel(A, \neg\Diamond p2)$. This leads to cancelling C's expected adoption beliefs. Acting as a non-cooperative but rational agent, A refuses to commit to $p2$. Alternatively, A may offer another value more preferable for him, i.e. performing a counter-offer when $Bel(A, Interest(A, \neg p2))$ but $Bel(A, Interest(A, p3)); Bel(A, CanDo(A, offer(p3)))$; $ConsidDo(A, offer(p3))$ where $p3$ stands for example for $offer(ISSUE = 1; VALUE = 1c)$.

5.7 Computing multidimensional states: evaluation

In order to assess the quality of the computed multimodal multidimensional information states, we compared the performance of the Debate Coach and Negotiation Task Agents with the human performance on the same tasks: Tutor and Experiencer.

To assess the system's formative real-time tutoring interventions, we conducted a series of experiments with human tutors evaluating debate performance. The output from these experiments has been used as simulated input for our tutoring system. Three human tutors provided feedback on the debaters' presentation, interactive and argumentative behaviour in real time by pressing a red button for negative feedback, e.g. 'talks too loud', 'talks too much', 'rude interruption', 'no evidence provided', 'unclear arguments', etc., and a green button for positive feedback. Two debate sessions were evaluated with a total duration of 22 minutes, consisting of 57 turns of 4 different speakers, and comprising 426 utterances. Time-stamped automatically generated tutoring interventions and those of a human Wizard were logged and compared. As can be observed in Table 5.10, human and system interventions differ a lot both quantitatively and qualitatively. The system generated about 50% more feedback messages, with a significantly higher portion of negative feedback than human tutors do. This does not mean that the system actions were wrong, however. Upon close inspection, the majority of them make perfect sense. Errors are attributed mostly due to imperfect interpretation of spoken trainee behaviour. Clearly, automatic natural language recognition and understanding are not ideal for many tasks. We found that important issues which are still largely open concern the amount, type and complexity of feedback which is appreciated most and considered useful. Thus, user-based evaluation and usability testing

[14] Consider also the definition of cooperative communicative behaviour proposed by Allwood et al., 2000. Communicative agents are considered cooperative at least in trying to recognise each other's goals, and the recognition of a goal may be sufficient reason for the participant to form the intention to act.

Aspect	Human Tutor		System	
	positive	negative	positive	negative
Presentation	0	26	14	58
Interaction	40	18	27	46
Structure	8	3	2	1
Totally	103		148	
Completely matched	40			

Table 5.10: Tutoring interventions generated by human tutors and by evaluated tutoring system.

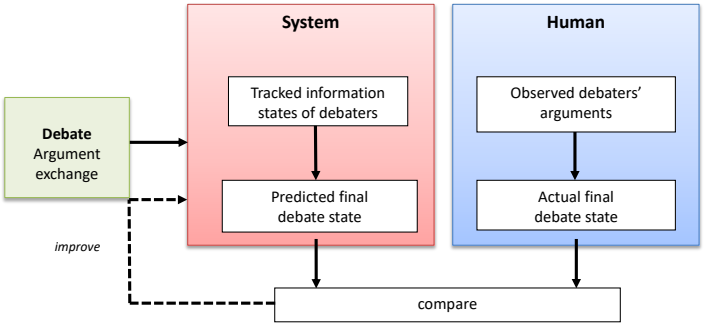

Figure 5.7: Evaluation model for the debate task success summative feedback.

are very important and are performed involving both trainees and tutors (see Chapter 6). From the evaluation with trainees insights are gained on what skills and what aspects are most important for them to master, and from the evaluation with tutors what type, amount and timing of interventions lead to the best learning outcome.

To assess the overall task success for summative feedback generation, the content of the system final information state (predicted outcome) was compared with the human tutors and human debaters' understanding of the final debate state (actual and assumed state). For this purpose, human tutors were asked to summarise (write down) the debate session by considering four debate rounds with respect to agreements achieved, i.e. arguments supported by the majority of debaters, and disagreements stated, i.e. arguments attacked by the the majority of debaters. Tutors were allowed to make notes during the debate session and replay the recorded debate session. The evaluation method is depicted in Figure 5.7.

Table 5.11: Example of System (S) predicted information state vs Human Tutor (HT) actual final information state . (pred.und = predicted understanding; und = understanding; pred.ad= predicted adoption; ad = adoption; pred.canc = predicted cancelling; canc = cancelling; Bel = believes; MBel = mutually believed; WBel = weakly believes)

source	System (S)	source	Human Tutor (HT)
pred.und	$Bel(S, MBel(\{S, D_1, D_3, D_4, D_6\},$ $Bel(\{D_1, D_3, D_4, D_6\}, arg_1)))$	und	$Bel(HT, MBel(\{HT, D_1, D_3, D_4, D_5, D_6\},$ $Bel(\{D_1, D_3, D_4, D_5, D_6\}, arg_1)))$
	$Bel(S, MBel(\{S, D_1, D_3, D_4, D_6\},$ $Want(\{D_1, D_3, D_4, D_6\},$ $Bel(S, arg_1))))$		$Bel(HT, MBel(\{HT, D_1, D_3, D_4, D_5, D_6\}$ $Want(\{D_1, D_3, D_4, D_5, D_6\},$ $Bel(HT, arg_1))))$
	$Bel(S, MBel(\{S, D_2, D_5\}, Bel(\{D_2, D_5\} \neg arg_1)))$		$Bel(HT, MBel(\{HT, D_2\}, Bel(D_2 \neg arg_1)))$
	$Bel(S, MBel(\{S, D_2, D_5\},$ $Want(\{D_2, D_5\}, Bel(S, \neg arg_1))))$		$Bel(HT, MBel(\{S, D_2\},$ $Want(D_2, Bel(HT, \neg arg_1))))$
	$Bel(S,$ $MBel(\{S, D_1, D_5\},$ $Bel(\{D_1, D_5\}, arg_2)))$		$Bel(HT,$ $MBel(\{HT, D_1, D_5\},$ $Bel(\{D_1, D_5\}, arg_2)))$
	$Bel(S, MBel(\{S, D_1, D_5\},$ $Want(\{D_1, D_5\}, Bel(S, arg_2))))$		$Bel(HT, MBel(\{HT, D_1, D_5\},$ $Want(\{D_1, D_5\}, Bel(HT, arg_2))))$
	$Bel(S, MBel(\{S, D_2, D_3, D_4, D_6\},$ $Bel(\{D_2, D_5, D_4, D_6\}, \neg arg_2)))$		$Bel(HT, MBel(\{HT, D_2, D_3, D_4, D_6\},$ $Bel(\{D_2, D_5, D_4, D_6\}, \neg arg_2)))$
	$Bel(S, MBel(\{S, D_2, D_5, D_4, D_6\},$ $Want(\{D_2, D_5, D_4, D_6\}, Bel(S, \neg arg_2))))$		$Bel(HT, MBel(\{HT, D_2, D_5, D_4, D_6\},$ $Want(\{D_2, D_5, D_4, D_6\}, Bel(S, \neg arg_2))))$
	$Bel(S, MBel(\{S, D_1, D_3, D_4, D_6\},$ $Bel(\{D_1, D_3, D_4, D_6\}, arg_3)))$		$Bel(HT, MBel(\{HT, D_1, D_2, D_3, D_4, D_5, D_6\},$ $Bel(\{D_1, D_2, D_3, D_4, D_5, D_6\}, arg_3)))$
	$Bel(S, MBel(\{S, D_1, D_3, D_4, D_6\},$ $Want(\{D_1, D_3, D_4, D_6\}, Bel(S, arg_3))))$		$Bel(HT, MBel(\{HT, D_1, D_2, D_3, D_4, D_5, D_6\},$ $Want(\{D_1, D_2, D_3, D_4, D_5, D_6\}, Bel(HT, arg_3))))$
	$Bel(S, MBel(\{S, D_2, D_5\}, Bel(\{D_2, D_5\}, \neg arg_3)))$		
	$Bel(S, MBel(\{S, D_2, D_5\},$ $Want(\{D_2, D_5\}, Bel(S, \neg arg_3))))$		
	$Bel(S, MBel(\{S, D_1, D_3, D_4, D_5, D_6\},$ $Bel(\{D_1, D_3, D_4, D_5, D_6\}, arg_4)))$		$Bel(HT, MBel(\{HT, D_1, D_3, D_4, D_5, D_6\},$ $Bel(\{D_1, D_3, D_4, D_5, D_6\}, arg_4)))$
	$Bel(S, MBel(\{S, D_1, D_3, D_4, D_5, D_6\},$ $Want(\{D_1, D_3, D_4, D_5, D_6\}, Bel(S, arg_4))))$		$Bel(HT, MBel(\{HT, D_1, D_3, D_4, D_5, D_6\},$ $Want(\{D_1, D_3, D_4, D_5, D_6\}, Bel(HT, arg_4))))$

source	System (S)	source	Human Tutor (HT)
	$Bel(S,MBel(\{S,D_2\},Bel(D_2,\neg arg_4)))$		$Bel(HT,MBel(\{HT,D_2\},Bel(D_2,\neg arg_4)))$
	$Bel(S,MBel(\{S,D_2\},$		$Bel(S,MBel(\{S,D_2\},$
	$Want(D_2,Bel(S,\neg arg_4))))$		$Want(D_2,Bel(HT,\neg arg_4))))$
	$Bel(S,MBel(\{S,D_1,D_2,D_3,D_4,D_5,D_6\},$		$Bel(HT,MBel(\{HT,D_1,D_2,D_3,D_4,D_5,D_6\},$
	$Bel(\{D_1,D_2,D_3,D_4,D_5,D_6\},arg_5)))$		$Bel(\{D_1,D_2,D_3,D_4,D_5,D_6\},arg_5)))$
	$Bel(S,MBel(\{S,D_1,D_2,D_3,D_4,D_5,D_6\},$		$Bel(HT,MBel(\{HT,D_2,D_7,D_{10},D_{14}\},$
	$Want(\{D_2,D_7,D_{10},D_{14}\},Bel(S,arg_5))))$		$Want(\{D_1,D_2,D_3,D_4,D_5,D_6\},Bel(HT,arg_5))))$
pred.ad	$Bel(S,MBel(\{S,D_1,D_3,D_4,D_6\},arg_1))$	ad	$Bel(HT,MBel(\{HT,D_1,D_3,D_4,D_5,D_6\},arg_1))$
	$Bel(S,MBel(\{S,D_1,D_5\},arg_2))$		$Bel(HT,MBel(\{HT,D_1,D_5\},arg_2))$
	$Bel(S,MBel(\{S,D_1,D_3,D_4,D_6\},arg_3))$		$Bel(HT,MBel(\{HT,D_1,D_2,D_3,D_4,D_5,D_6\},arg_3))$
	$Bel(S,MBel(\{S,D_1,D_3,D_4,D_5,D_6\},arg_4))$		$Bel(HT,MBel(\{HT,D_1,D_3,D_4,D_5,D_6\},arg_4))$
	$Bel(S,MBel(\{S,D_1,D_2,D_3,D_4,D_5,D_6\},arg_5)$		$Bel(HT,MBel(\{HT,D_1,D_2,D_3,D_4,D_5,D_6\},arg_5))$
pred. canc	$Bel(S,MBel(\{S,D_2,D_5\},$	canc	$Bel(HT,MBel(\{HT,D_2\},$
	$WBel(S,Bel(\{D_2,D_5\},\neg arg_1))))$		$WBel(HT,Bel(D_2,\neg arg_1))))$
	$Bel(S,MBel(\{S,D_2,D_5\},$		$Bel(HT,MBel(\{HT,D_2,D_5\},$
	$WBel(S,Bel(\{D_2,D_5,D_4,D_6\},\neg arg_2))))$		$WBel(HT,Bel(\{D_2,D_5,D_4,D_6\},\neg arg_2))))$
	$Bel(S,MBel(\{S,D_2,D_5\},$		
	$WBel(S,Bel(\{D_2,D_5\},\neg arg_3))))$		
	$Bel(S,MBel(\{S,D_2\},$		$Bel(HT,MBel(\{HT,D_2\},$
	$WBel(S,Bel(D_2,\neg arg_4))))$		$WBel(HT,Bel(D_2,\neg arg_4))))$

Evaluation criteria	Human-human	Human-computer
Number of dialogues	25 (5808)	185 (NA)
Mean dialogue duration (in turns per dialogue)	23 (6.6)	40 (NA)
Agreements (%)	78 (80.1)	66 (57.2)
Pareto optimal (%)	61 (76.9)	60 (82.4)
Negative deal (%)	21 (NA)	16 (NA)
Cooperativeness rate (%)	39 (NA)	51 (NA)

Table 5.12: Comparison of human-human and human-agent negotiation behaviour. Adopted from Petukhova et al. (2017). In brackets the best results reported by Lewis et al. (2017) for comparison. *NA* stands for not applicable, i.e. not measured.

We compute the S beliefs by applying the analysis exemplified in (10) to a summary given by a human tutor in (11). For S we compute the list of predicted beliefs resulting from understanding, grounding and the arguments supported by a 'winning' majority. The *predicted* final system information state and computed tutor's *actual* states are compared. Table 5.11 presents the predicted and actual final information states. The representation of expected understanding effects has been omitted both for the system and tutor, since they are identical. The propositions arg_1 to arg_5 stand for arguments.[15]

As we can observe, the predicted information state differs slightly from the actual information state, but not significantly. The human tutor interpreted that arg_1 was attacked only by D_2 and considered the position of D_5 as supporting arg_1. Human tutors also did not find evidence for (Dis-)Agreement acts with arg_3 and considered arg_3 as adopted and not cancelled. These differences can be attributed to the imperfect system understanding, especially negations present a problem.

Evaluating the DM performance in the negotiation setting, we compared it with human performance on the number of agreements reached, the ability to find Pareto optimal outcomes, the degree of cooperativeness, and negative outcomes. For this evaluation, 28 sessions involving 28 participants aged 25-45 (all professional politicians or governmental workers) were analysed.

It was found that participants reached a lower number of agreements when negotiating with the system than when negotiating with each other, 66% vs 78%. Participants made a similar number of Pareto optimal agreements (about 60%). Human participants show a higher level of cooperativity when interacting with the system, i.e. 51% of the actions are perceived as cooperative. This may mean that humans were more competitive when interacting with each other. A lower number of negative deals was observed for human-agent pairs, 21% vs 16%. Users perceived their interaction with the system as effective when they managed to complete their tasks successfully reaching Pareto optimal agreements by performing cooperative actions but avoiding excessive concessions. Our results differ from those reported in Lewis et al. (2017) for both the human-human and the human-agent setting, see Table 5.12. However, as noticed above, due to differences in tasks, scenario and interactive setting it is hard to draw clear comparative conclusions. Nevertheless, we can conclude that the implemented NA is capable of making decisions and performing actions

[15]For the sake of simplicity we do not spell out the semantic content of the arguments and leave out evidence links here.

similar to those of humans. No significant differences in this respect were observed between human-human and human-system interactions.

5.8 Summary

We have presented an approach to dialogue management that integrates a cognitive task agent able to reason about the goals and strategies of human partners, and to successfully engage in a negotiation task. This agent leverages established cognitive theories, namely ACT-R and instance-based learning, to generate plausible, flexible behaviour in this complex setting. We also argued that separate modelling of task related and dialogue control actions is beneficial for current and future dialogue system designs. The implementation introduced a theoretical novelty in instance-based learning for Theory of Mind skills and integrating this in the dialogue management of a tutoring system. The Cognitive Task Agent used instance knowledge not only to determine its own actions, but also to interpret the human user's actions, allowing it to adjust its behaviour to its mental image of the user. This work was successful: human participants who took part in evaluation experiments were not able to discern human users from simulated task agents (see also [Stevens et al., 2016b]), and an agent using Theory of Mind prompted users to use that themselves. Our evaluation results suggest that the dialogue system with the integrated cognitive agent technology delivers plausible negotiation behaviour leading to reasonable user acceptance and satisfaction.

The work presented here has certain limitations. Instances in the instance-based learning model, slots, values and preferences for both partners, were largely pre-programmed, which limits its general applicability. In the future, the agent will learn from real human-human dialogues, e.g. extract negotiation issues and values, and assess their importance. We will also enable the collaborative creation and real-time interactive correction, (re-)training and generation of agents by domain experts and target users. We aim to design authoring tools supporting agent learning and re-training across different situations.

Furthermore, we successfully integrated cognitive, interaction and learning models into a baseline proof-of-concept system. More research is needed on the connections between the cognitive models and the interaction and learning models, and overall mechanisms need to be further specified that underlie communication strategies depending on information about the current state of the task, participants' (learning) goals, participant's affected state, and the interactive situation/environment. Negotiation is more than the exchange of offers, decision making or problem-solving; it involves a wide range of aspects related to feelings, emotions, social status, power, and interpersonal relations, context and situation awareness. For instance, tentative cooperative actions can engender a positive reaction and build trust over time, while social barriers can trigger interactive processes that often lead to bad communication, polarisation and conflict escalation (Sebenius, 2007). Such dynamics may be observed in negotiations involving participants of different genders, races, or cultures (Nouri et al., 2017). Aspects related to social and interpersonal relations like dominance, power, politeness, emotions and attitudes deserve substantially more attention. We aim to

advance our knowledge on social cognition that models human capabilities.

Finally, recent advances in digital technologies open new possibilities for us to interact with our environment, as well as for our environment to interact with us. Everyday artefacts which previously did not seem aware of the environment at all are turning into smart devices and smart toys with sensing, tracking or alerting capabilities. This offers many new ways for real-time interaction with highly relevant, social and context-aware agents in multimodal multisensory environments which, in turn, enables designing rich immersive interactive experiences.

Chapter 6

Application: Virtual Coaching

This chapter discusses two interactive tutoring applications - Virtual Negotiation and Virtual Debate Coaches - with the main goal to evaluate the proposed approach in terms of technical system performance and user acceptance. For this, the system was integrated following the principles of multimodal dialogue system architectures including core components related to recognition, interpretation, management and generation. The chapter describes the technical architecture and provide details for each integrated module and report testing results. The proof of concept systems were evaluated in trainee-based settings.

Introduction

International research has so far produced knowledge-based and statistical multimodal dialogue systems capable of interacting with structured data bases, e.g., train time tables, hotel and restaurants reservations. Such systems typically exhibit reactive behaviour. Single or pre-defined strategies are pursued. System architectures are typically pipelined where output of one module serves as input for the following module. What is more important, such systems lack the notion of reflection about their own behaviour, i.e. metacognitive capabilities. In other words, explicit reasoning mechanisms are required to assess why a particular solution worked or not, and manipulate the task representation accordingly.

The designed proof-of-concept dialogue systems, particularly their adaptive and flexible Dialogue Management components, integrate cognitive task agents with metacognitive skills including exploring and reasoning about task-related behaviour, adapting and training

Section 6.1 summarises the efforts of the technical team of the Metalogue project performed when designing and implementing individual system modules, the reported results are appropriately cited. The overall integration approach, in particular the inter-module communication design based on ZeroMQ message passing protocols, the dialogue act recognition module integration, and the development and integration of the Dialogue Manager reported here were developed by me. Section 6.2 is based on Malchanau et al.(2018a), for which I performed the research in close collaboration with my co-authors. The evaluation experiments were conducted with the assistance of Dimitris Koryzis, Hellenic Parliament, and Dimitris Spiliotopoulos, University of Peloponnese, Greece; the interpretation of the results is mine.

to adapt the behavior at the level of setting goals, choosing appropriate strategies and monitoring progress, predicting and improving targeted outcomes. The reference architecture for dialogue system that incorporates metacognitive processes and supports the development of metacognitive skills by the human trainee was implemented and evaluated.

Multimodal dialogue is proven to be the most natural form of interactive learning by offering a mode of interaction that has certain similarities with human natural communication, using input and output modalities that people normally employ in the learning process, including speech, gesture, facial expressions, touch and point. The educational dialogue and tutoring interventions provide useful constraints and a dialogue context. Despite the remarkable progress booked, absolutely free multimodal natural interaction is still not feasible due to certain technological limitations, e.g. imperfect Automatic Speech Recognition or visual movement tracking. For the tutoring task fully free interaction may, however, not be required, since tutoring interaction is never a fully free and fully natural dialogue. There is a tutor who proposes certain restrictions and guides the trainee through the learning process. Learners, even novice ones, are familiar with tutoring settings and procedures, and do not experience such restrictions as something unnatural or artificial. It is, therefore, a reasonable strategy to have pre-defined evaluation scenarios for the whole or parts of the interaction. This also assures that the system's performance is as robust and as natural as possible. The other promising way to achieve reliable system performance is the restriction with respect to the modalities used. For example, the speech modality is restricted or alternatives are offered, e.g. in some modes used only by the system, and when also used by the user then supported by control/selection actions using a graphical user interface, or only for certain parts of the interaction. Additionally, the interactive strategies and tactics are clearly described in the logic of educational dialogue situations which enables an adequate evaluation of dialogue system performance, since we would have a clear goal as a measure for dialogue effectiveness.

The chapter is structured as follows. Section 6.1 outlines the system technical architecture, provides important details for key modules and inter-module communication design. Section 6.2, presents the scenarios and results of user-based evaluation experiments assessing the relative contribution of various factors to the overall usability of a dialogue system. Subjective perception of effectiveness, efficiency and satisfaction were correlated with various objective performance metrics, e.g. number of (in)appropriate system responses, recovery strategies, and interaction pace.

6.1 System architecture

Considering the use cases specified in Section 4.3, we defined the capabilities that the Virtual Coach for metacognitive skills training should have. Capabilities are realised through specific modules and include audio-visual sensing, multimodal behaviour analysis, dialogue management with the integrated cognitive task agents, and behaviour realisation and rendering. For sensing and multimodal behaviour realisation voice input/output technology was combined with additional multimodal input/output capabilities integrating Kinect 3D

tracking, touch screen control gestures and full body humain-like avatar. A modular, fully integrated and open architecture was designed. The developed system is not limited to one application domain, use case or technical solution: the core components are applicable outside the negotiation and/or debate domain (e.g. medical, human resource management, etc.), the proposed architecture can be extended to other use cases (e.g. custom support management training) and to novel processing algorithms and emerging HCI and AI technologies. Other modern devices and sensors (e.g. GPS positioning, web cameras, eye-trackers, biometric sensors, etc.) could be considered for future extensions. The overall architecture, together with most of its components, is represented schematically in Figure 6.1.

As proofs of concept, and for assessing the potential value of the integration of a cognitive task agents into a dialogue manager, we designed the Virtual Negotiation Coach (VNC) and Virtual Debate Coach (VDC), interactive tutoring systems with the functionality described in the scenario for data collection (Section 4.4). The VNC and VDC get a multimodal signal, recognise and interpret it, identifies relevant actions and generates multimodal actions, i.e. speech and gestures of a virtual negotiator represented by the full body avatar, positive and negative visual feedback for tutoring in debate and negotiation setting, e.g. as incorporated in the Presentation Trainer displayed in Figure 5.6, negotiation actions are also displayed by the graphical interface representing preference cards as depicted in Figure 4.9. We further describe the key processing modules and communication between them.

6.1.1 Multimodal input recognition

Speech is one of the main modalities in human-computer multimodal dialogue systems. Significant progress has been made in Automatic Speech Recognition (ASR), which has often been considered as the main obstacle in making natural language dialogue interfaces robust. For more than 30 years, ASR has been dominated by statistical modelling schemes, such as Hidden Markov Model (HMM)/Gaussian Mixture Models (GMMs) (Gales and Young, 2007) and n-gram language models (Rosenfeld, 2000). This has resulted in well-established core algorithms that are largely language independent and have proven to work reasonably well for a large variety of languages. Statistical models require however large amounts of transcribed training data for robust and reliable parameter estimation.

In our technical setup, speech signals are recorded from multiple sources, such as wearable microphones (PinMic lapel microphones), headsets (Sennheiser PC 3 headsets) for each dialogue participant, and an all-around microphone (Tascam Dr-40 recorder) placed between participants. The recordings were performed in the following setting: sample rate (48KHz), sample size (16-bit), sample format (linear PCM) with stereo channel which was later converted to mono. The Kaldi-based ASR component incorporates acoustic and language models developed using various available data sources: the Wall Street Journal WSJ0

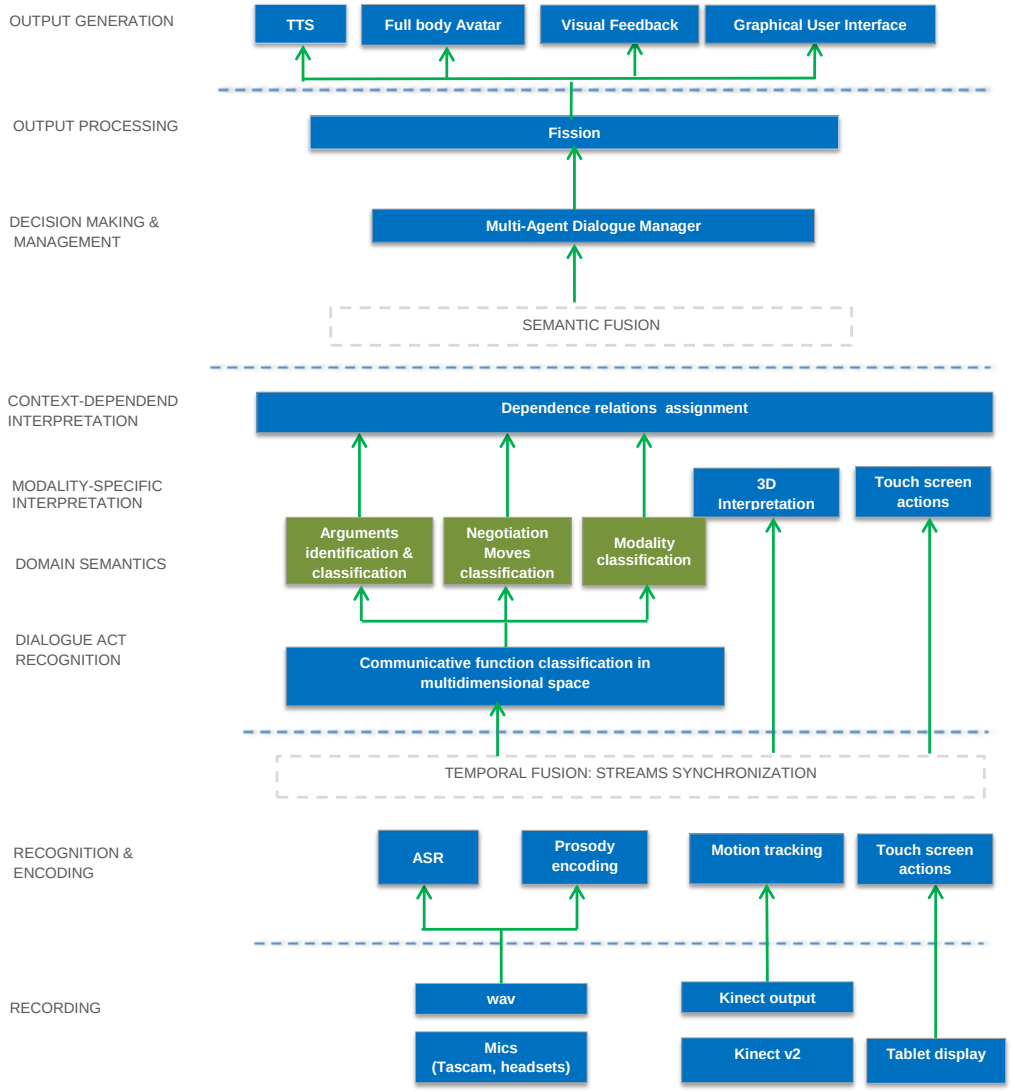

Figure 6.1: Open architecture of the general Virtual Coaching system. From bottom to top, signals are received through input devices, further recognised by tailored processing modules. After interpretation concerned with events, arguments, modality and communicative functions classification, semantic representations from different modalities and modules are fused as Dialogue acts. Fused dialogue act information is passed to the Dialogue Manager for context model update and next action generation. The generated system response is rendered or 'fissed' in different output modalities.

.

corpus[1], HUB4 News Broadcast data[2], the VoxForge corpus[3], the LibriSpeech corpus[4] and AMI project data[5]. In total, about 759 hours of data has been used to train an acoustic model. The collected in-domain negotiation data is used as language model adaptation. The background language model is based on a combination of different corpora, like the approach taken to train the acoustic model. The ASR performance is measured at 34.4% Word Error Rate (WER), see [Singh et al., 2017][6]. The speech signals serve as input for two types of processing: Automatic Speech Recognition, which leads to lexical, syntactic, and semantic analysis, and prosodic analysis concerned with voice quality, fluency, stress and intonation. For the former use, the ASR outputs a single best word sequence without any scores. Prosodic properties were computed automatically using PRAAT [Boersma and Weenink, 2009] such as minimum, maximum, mean, and standard deviation of *pitch*, *energy*, *voicing* and speaking rate.[7]

The current state of the technology in markerless motion capture is mature enough in order to boost the research on the recognition of action units using off-the-shelf and affordable equipment, such as webcams for facial expression tracking or Intel Real Sense technology (Dornaika and Davoine, 2006; Dornaika and Raducanu, 2009), and depth sensing devices for full-body tracking such as Microsoft Kinect sensors (Shotton et al., 2011). Today, motion capture technology allows measuring human body motions precisely. We are facing an open problem though: how can intelligent systems use motion capture data? Automatic recognition of human behaviour is one of the common research challenges in human motion analysis. Motion capture data is not easy to recognise in general. The data usually have a large amount of dimensions to record on XYZ trajectory of major parts of a human body. High dimensionality makes recognition difficult. For our system the depth sensing camera - Microsoft Kinect 3D sensor - was used. Additional Track software was designed to track visible motions in real-time and to smooth the tracking data for further robust processing and interpretation.

6.1.2 Semantic processing

The ASR output is used for the interpretation of the participants' verbal contributions. Interpretation of dialogue behaviour is primarily based on the recognition of the speaker's intentions encoded in the communicative function. Additionally, 'dimensions' have been used to classify communicative functions in multidimensional space assigning shallow semantic

[1] `https://catalog.ldc.upenn.edu/ldc93s6a`
[2] `https://catalog.ldc.upenn.edu/ldc98s71`
[3] `http://www.voxforge.org/`
[4] `http://www.openslr.org/12/`
[5] `http://groups.inf.ed.ac.uk/ami/corpus/`
[6] It should be noticed that the ASR performance has been measured when interacting with non-native English speakers, who significantly varied in language skills level and speech fluency, some having a rather strong Greek accent.
[7] We computed both raw and normalised versions of these features. Speaker-normalised features were obtained by computing z-scores (z = (X-mean)/standard deviation) for the feature, where mean and standard deviation were calculated from all functional segments produced by the same speaker in the debate session. We also used normalisations by the first speaker turn and by prior speaker turn.

Type of gesture		Relative frequency (in %)	Proportion of total 7074 frames (in %)
Beats	all categories	59.55	27.04
	prominence intensifier	69.76	68.90
	new topic/theme marker	3.26	3.45
	meta-discursive act marker	17.67	16.36
	phrase/boundary marker	9.31	11.29
Adaptors		14.96	18.80
Iconic		2.22	1.37
Deictic		2.22	1.84
Emblem		0.55	0.24
No visible gesture event		20.50	50.7

Table 6.1: Detected gesture events distribution in terms of their relative frequency (in%) and proportion of frames (in %). Adopted from Petukhova et al., 2017

interpretations (semantic content type), see [Bunt, 2011, ISO, 2012] and [Petukhova, 2011]. For this purpose, various machine learning techniques have been applied, such as Support Vector Machine (SVM, Boser et al., 1992), Logistic Regression (Yu et al., 2011), AdaBoost (Zhu et al., 2009), and the Linear Support Vector Classifier (Vapnik, 2013). F-scores ranging between 0.83 and 0.86 were obtained, which corresponds to the state-of-the-art performance, see [Amanova et al., 2016]. The incremental token- and chunk-based communicative functions CRF-classifiers showed a performance of .80 F-scores on average, see [Ebhotemhen et al., 2017]. After extensive testing, a non-incremental SVM-based classifier has been integrated in to the Virtual Coaching system.

Domain-specific interpretation is concerned with the classification of Argumentative Discourse Units (ADUs) for the VDC system and (modalised) Negotiation Moves for the VNC system. ADUs are identified based on the recognised discourse relations (Petukhova et al., 2017b). The SVM-based classifier yielded F-scores of 0.54 on a coarse 3-class task (Contingency, Evidence, No-Relation) and 0.46 on a fine-grained 7-class task (Justification, Reason, Motivation, Exemplification, Explanation, Exception and No-Relation). Negotiation moves specify events and their arguments represented as $NegotiationMove(ISSUE; VALUE)$. Conditional Random Field models for sequence learning (CRF, Lafferty et al., 2001) were trained to predict three types of classes (move, issue and value) and their boundaries in ASR n-best strings: negotiation move, issue, preference value. A ten-fold cross-validation using 5000 words of transcribed speech from the negotiation domain yielded an F-score of 0.7 on average. The Support Vector Machine [Vapnik, 2013] modality classifiers were trained to classify expressions of necessity, preferences, acquiescence and abilities, and showed accuracies in the range between 73.3 and 82.6% [Petukhova et al., 2017a]. The output of this module is the interpretation of a modalised negotiation move, e.g. stating preference is represented as $\Box offer(ISSUE = X; VALUE = Y)$.

Kinect tracked data is used to detect hand/arm co-speech gestures[8] and their types, see Table 6.1. SVM and Gradient Boosting [Friedman, 2002] classifiers were trained and achieved F-scores of 0.72 [Petukhova et al., 2017b]. The motion interpretation component related to hand/arms position detection of the designed Presentation Trainer (Van Rosmalen et al., 2015; Schneider et al., 2015a) is integrated into the VDC system.

[8]Co-speech gestures are visible hand/arm movements produced alongside speech and are interpretable only through their semantic relation to the synchronous speech content.

The system includes a Fusion component, which combines the modality-specific analyses into a fused representation of participant's multimodal actions in terms of dialogue acts. Prior to this, information from the Linguistic Context related to the dialogue history has been used to ensure context-dependent interpretation of dialogue acts. Subsequently, at the semantic fusion level, verbal, prosodic and motion tracking information is combined to obtain complete context-dependent multimodal dialogue act interpretations. For instance, prosodic and motion tracking information has been combined to interpret the status of information conveyed in an argument. Exploiting the fact that pitch-accented tokens often coincide with focus, topic and contrast, and if accompanied by a beat gesture are perceived as even more prominent, we identified 95% of all beat gesture events produced around intensity peaks. The fusion module also incorporates an SVM-based classifier that operates on prosodic and motion features, and predicts the persuasiveness level of an argument with an accuracy of 71% [Petukhova et al., 2017c].

6.1.3 Dialogue management

Given the system's understanding of the trainee's behaviour, the core VDC and VNC task is to perform tutoring interventions to inform the trainee of a mistake or to propose corrections, or to provide positive feedback. The performance on this task requires immediate real-time feedback, often called 'in-action' feedback (Schön, 1983) on the aspects and criteria defined in Table 5.1, Section 5.2.

For tutoring in-action feedback on the presentation aspects, the participant's postures are detected. The information state is continuously updated using dialogue acts generated by the Fusion module. When certain inappropriate behaviour (postures or prosody types) are detected for more than a pre-set time span (1000ms), the DM plans a negative feedback inform or correction act, and sends it to the Fission module for generation. Subsequently, when a participant's posture was corrected for a pre-set time span (500ms) an inform about positive execution of an action is generated. In case there are several posture mistakes, the system feedback about these events is provided sequentially.

For tutoring in-action feedback on the participant's negotiation strategy, and feedback on chosen/trained negotiation strategy, e.g. either 'cooperative' or 'aggressive' is generated.

The VNC system is designed with the ability to 'participate' as a full-fledged partner in multi-issue bargaining dialogues, namely, in the role of a Small Business Representative. Thus, the system generates negotiation actions, dialogue acts with negotiation moves as semantic content.

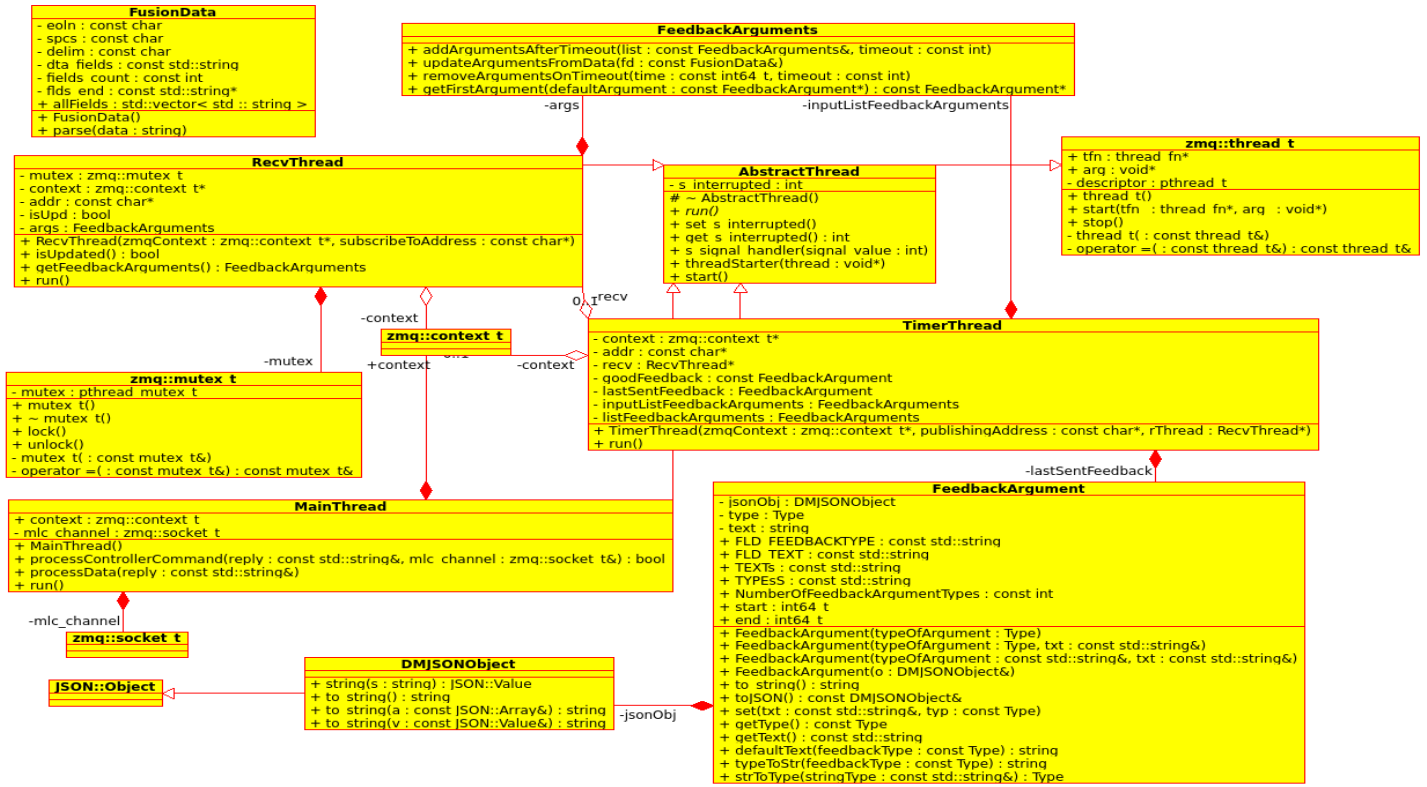

Figure 6.2: The UML class diagram of the Dialogue Manager integrated in the Virtual Debate Coach.

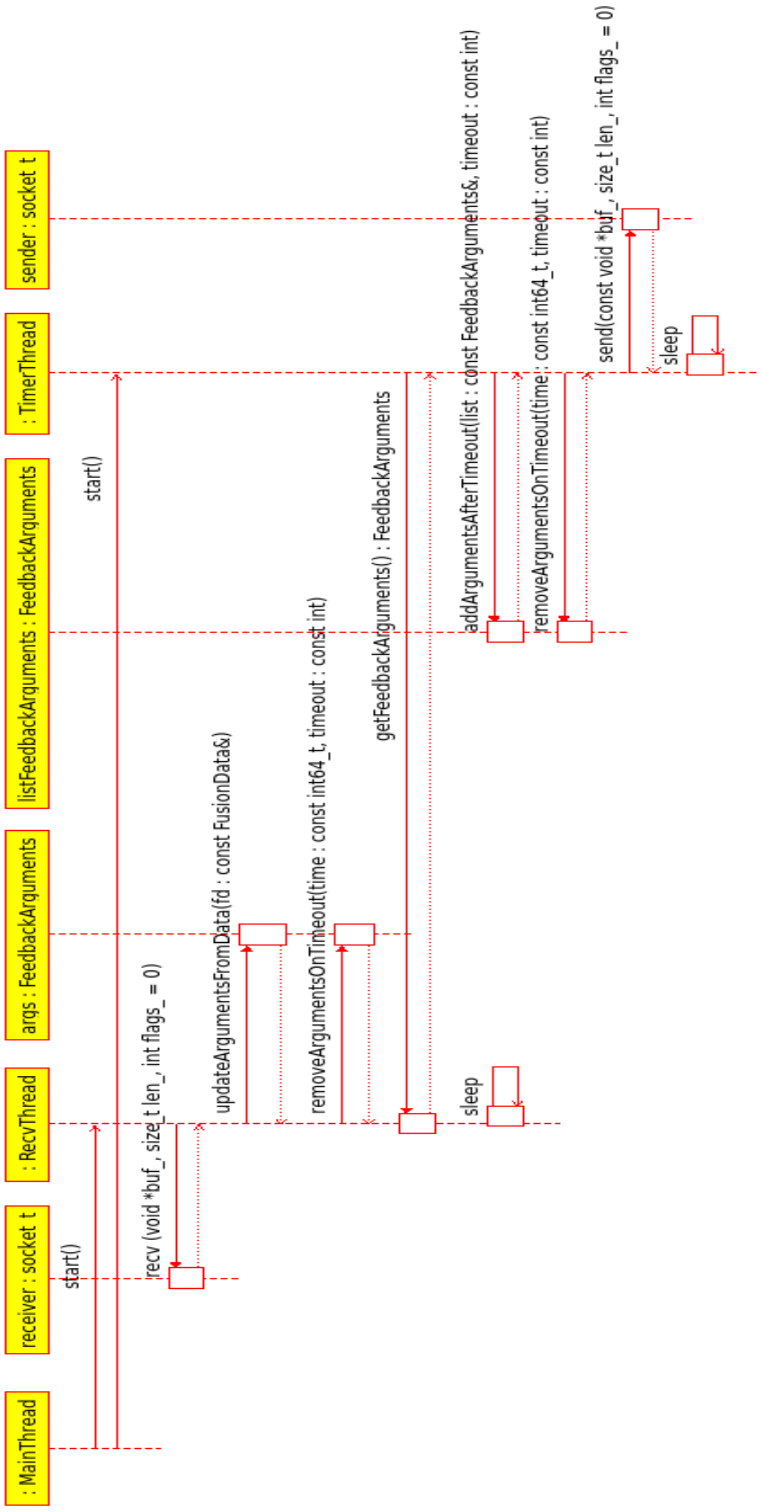

Figure 6.3: UML sequence diagram of the Dialogue Manager integrated in the Virtual Debate Coach.

Data is sent as <frame> element. A
tt ¡frame¿ may include sub-elements (multiple occurrences) for *primary data* (token, sound, tracking), *segmented data* in the form of functional and feedback segments, and *dialogue acts*. Semantic content of dialogue acts (i.e. negotiation moves) is sent via content sub-elements of a <dialogueAct> element. For example:

```
<frame>
  <fs id="fs1" sender="#p1"/>
  <dialogueAct id="da1" target="#fs1" sender="#p1"
               addressee="#p2" dimension="task"
               communicativeFunction="inform"
               sentiment="positive">
    <negotiationSemantics>
      <negotiationMove xml:id="nm1" type="offer">
        <arg><value>1a</value></arg>
      </negotiationMove>
    </negotiationSemantics>
  </dialogueAct>
</frame>
```

The Dialogue Manager is written in C++ and compiles under Linux and Windows operating systems. Microsoft Visual Studio solution is available for the Windows version and an Eclipse project file is available for the Linux version. The code of the module contains conditional macro definitions that enable compilation for multiple OS. UML class and sequence diagrams are presented in Figure 6.2 and 6.3 respectively.

6.1.4 Multimodal output rendering

Given the dialogue acts provided by the Dialogue Manager, the Fission module generates system responses, splitting content into different modalities, such as Avatar[9], Voice (TTS[10]) and visual feedback for tutoring interventions. For the former, the results of the DM planning phase is a set of actions that has to be rendered by the system using a combination of generated speech output and primitives available from libraries for, e.g. gestures, facial expressions and body postures that can be performed by the full-body avatar. The latter includes a representation of the negotiators' current cooperativeness, visualised by happy and sad face emoticons. Additionally, the trainee has a choice to select options using a graphical interface as depicted in Figure 4.9. As task progress support, both partners' negotiation moves and agreements are visualised with red (system) and green arrows (user).

At the end of each debate session, summative feedback is generated summarising the number of arguments, hesitations, interruptions, editing expressions, etc.

[9] Commercial software of Charamel GmbH has been used, see [Reinecke, 2003]

[10] Vocalizer of Nuance, http://www.nuance.com/for-business/text-to-speech/vocalizer/index.htm, was integrated.

6.1.5 Inter-module communication

The communication infrastructure, implemented as a set of libraries for C++ (both Windows and Linux versions), Java and Python provides several functionalities for worker modules:

- passing messages between modules, based on chosen patterns of communication;
- dynamic or static configuration (acquiring, storing, providing);
- watch-dog functions (heart-beat, start/restart);
- logging functions;
- time and synchronisation services;
- start/stop and other control services;
- data format translation (e.g. translating between XML data representations and module internal binary structures)

Additional requirements considered in the design are concerned with managing of available computational, memory and storage resources by each individual module. For example, in a case when the history of a particular channel(s) is required, module developers make sure they have enough memory and/or disk space for the amounts of information they want to keep. Secondly, detection and reporting of duplicate registration requests from different modules is required. This also includes cases when modules request to be a publisher for a channel that is being already requested by another module.

Passing messages between modules is built on top of asynchronous messages-processing transportation by the ZeroMQ library[11]. ZeroMQ provides support for several communication patterns. Some of the ZeroMQ patterns that were used are:

- Request-reply: a distribution pattern which remotely connects a set of clients to a set of services;
- Publish-subscribe: a data distribution pattern which connects a set of publishers to a set of subscribers;
- Pipeline: a parallel task distribution and collection pattern which connects nodes that can be multiple steps and loops; and
- Exclusive pair, which connects two sockets exclusively.

In general data was sent between modules as text(strings). When a standard representation of data was available or agreed upon between module developers it was possible to utilise XML parsing libraries. Whenever appropriate XSD files that describe XML data were available a partially automatic generation of code for dealing with data simplified development of the modules.

[11] ZeroMQ (0*MQ*) stands for "Zero latency Message Queuing". It achieves a concurrency framework in scalable distributed applications, based on a networking library designed in socket-style API.

6.2 Multimodal system evaluation

As a part of the interactive application design, evaluations are performed in order to assess the success of the developed solutions. Evaluation results serve to inform designers about the system's functional and non-functional deficiencies.

Dialogue systems are often exposed to user-based evaluation. This is commonly done by asking users to fill in a questionnaire after interacting with the system. It is still largely an open question which parameters should be taken into account when designing a satisfaction questionnaire, and which of these may correlate well with user satisfaction. Qualitative and quantitative measures are often automatically computed from test interactions with real or simulated users. Most existing evaluation metrics are designed for task-oriented information seeking spoken dialogue systems and do not apply well to complex multimodal interactions.

Several dialogue system evaluation approaches have been proposed in the past. PARADISE, one of the most widely-used evaluation models [Walker et al., 1997], aims at predicting user global satisfaction given a set of parameters related to task success and dialogue costs. Satisfaction is calculated as the arithmetic mean of nine user judgments on different quality aspects rated on 5-point Likert scales. Subsequently, the relation between task success and dialogue cost parameters and the mean human judgment is estimated by means of a multivariate linear regression analysis.

Another approach is to evaluate a dialogue system on the basis of test interactions substituting human users by computer agents that emulate user behaviour, see e.g. López-Cózar et al., 2006. The various types of users and system factors can be systematically manipulated, e.g. using interactive, dialogue task and error recovery strategies.

As for system performance metrics and interaction parameters, several sets have been recommended for spoken dialogue system evaluation, ranging from 7 parameters as defined in [Fraser, 1998] to 52 in [Möller, 2004] related to the entire dialogue (duration, response delay, number of turns), to meta-communication strategies (number of help requests, correction turns), to the system's cooperativity (contextual appropriateness of system utterances), to the task which can be carried out with the help of the system (task success, solution quality), as well as to the speech input performance of the system (word error rate, understanding error rate).

When evaluating an interactive application, real user judgments provide valuable insights into how well the application meets user expectations and needs. One of the methods to measure users' attitudes is to observe their behaviour and establish links between their emotions and actions [Kooijmans et al., 2007]. Given the current technical possibilities, the tracking and analysis of large amounts of logged user-generated multimodal data has become feasible [Linek et al., 2008]. For instance, gaze re-direction, body movements, facial muscle contraction, skin conductivity and heart rate variance may serve as a source of information for analysing a user's affective state.

The most common practice is to solicit user judgments on different system quality aspects with the help of a questionnaire. The absence of standard questionnaires for dialogue systems evaluation makes it difficult to compare the results from different studies, and the various existing questionnaires exhibit great differences:

- The PARADISE questionnaire has nine user satisfaction related questions (Walker et al., 2000).
- The Subjective Assessment of Speech System Interfaces (SASSI) questionnaire contains 44 statements rated by respondents on 7-point Likert scales (Hone and Graham, 2001).
- The Godspeed questionnaire comprises 24 bipolar adjective pairs (e.g. fake-natural, inert-interactive, etc.) related to (1) anthropomorphism, (2) animacy, (3) likeability, (4) perceived intelligence and (5) perceived safety to evaluate human-robot interactions on 5-point Likert scales (Bartneck et al., 2009).
- The REVU (Report on the Enjoyment, Value, and Usability) questionnaire was developed to evaluate interactive tutoring applications and comprises 53 statements rated on 5-point Likert scales divided into three parts: OVERALL, NL (Natural Language), and IT (Intelligent Tutor) (Dzikovska et al., 2011).
- The Questionnaire for User Interface Satisfaction (QUIS[12], Chin et al., 1988) measures satisfaction related to (1) overall user reaction, (2) screen, (3) terminology and system information, (4) learnability, (5) system capabilities, (6) technical manuals and on-line help, (7) on-line tutorials, (8) multimedia, (9) teleconferencing, and (10) software installation. A short 6-dimensional form contains 41 statements rated on 9-point Likert scales, a long one has 122 ratings used for diagnostic situations.

The QUIS questionnaire is widely used and is considered as de-facto standard for user satisfaction assessment when performing usability studies. The QUIS forms can be customised by selecting evaluation aspects relevant for a specific application and use case, as we will show in the next sections when evaluating a multimodal dialogue system. Thus, we propose to assess multimodal dialogue system performance by relating various performance metrics, interaction parameters, and subjective perception of *usability* factors as defined by the ISO 9241-11 and ISO/IEC 9126-4 standards. This enables usability quantification in a meaningful and systematic way.

6.2.1 Usability definition

It is a common practice to evaluate an interactive system and its interface using a number of observable and quantifiable metrics for effectiveness, efficiency and satisfaction - see the ISO 9241-11 and ISO/IEC 9126-4 standards.

Task completion and the accuracy with which users achieve their goals are associated with the system's *effectiveness*. Task completion is calculated as the proportion of successfully completed tasks given the total number of tasks. To measure success of information retrieval tasks in information seeking dialogues, Attribute Value Matrix (AVM) metrics are used as proposed in PARADISE. In tutoring interactive applications, the task completion rate will depend on the system's ability to provide meaningful feedback (Dzikovska et al.,

[12] Version 7.0 is available `http://www.lap.umd.edu/QUIS/index.html`

2011). In the next section we will define effectiveness metrics for our negotiation training use case.

Efficiency is associated with the effort that users spend to perform specified tasks and is often correlated with temporal and duration properties of the interaction, e.g. number of turns, pace, reaction times, etc. Measures of efficiency associated with user's cognitive costs relate to [Dix, 2009]:

- *robustness*, referring to the level of support provided to the user in determining achievement and assessment of goals; is related to observability, recoverability, responsiveness and task conformance;
- *learnability*, referring to the ease with which new users can begin effective interaction and then to attain a maximal level of performance; is related to predictability, familiarity and consistency; and
- *flexibility*, referring to the multiplicity of ways in which the user and the system can communicate; is related to initiative, task substitutivity and customisability.

Satisfaction is concerned with user attitudes associated with the product use. Satisfaction is measured at the task and test levels. Popular post-task questionnaires are After-Scenario Questionnaire (ASQ, [Lewis, 1991]), NASA Task Load Index (TLX)[13] and Single Ease Question (SEQ)[14]. Satisfaction at the test level serves to measure users' impression of the overall ease of use of the system being tested.

In order to develop a reliable questionnaire for assessing user perception of a multimodal dialogue system usability we conducted an online study. QUIS 7.0 served as the basis for respondents to make their selection of aspects they think are important for them when evaluating a multimodal dialogue system. QUIS provides a useful decomposition of the usability concept into several dimensions (factors), enabling a clear mapping of system performance to distinctive usability perception aspects, with the advantage of being able to assess the impact of different items on usability perception instead of simply summing up or averaging to compute an overall satisfaction score (as e.g. in PARADISE or with the System Usability Scale, SUS [Brooke et al., 1996]). Adapting the QUIS questionnaire for the purposes of multimodal dialogue system evaluation, we considered factors assessed by the SAASI and Godspeed questionnaires. Previous studies showed that evaluative adjectives, bipolar adjective pairs and specific evaluative statements appeared to be more accurate than global satisfaction questions and were the most preferred forms for respondents [Chin et al., 1988, Root and Draper, 1983]. In our study, 36 evaluative adjectives, 40 bipolar adjective pairs, and 34 evaluative statements were ranked on 5-point Likert scales by 73 respondents, from which 69.6% considered themselves as dialogue researchers or related, and all respondents used dialogue systems at least once in their life. The study showed that important aspects related to user satisfaction are concerned with *task completion, task quality, robustness, learnability, flexibility, likeability, ease of use* and *usefulness/value* of

[13] `https://humansystems.arc.nasa.gov/groups/TLX/`

[14] A 7-point rated question to assess how difficult users find a task, see `https://measuringu.com/single-question/`

the application. We adopted the QUIS 7.0 structure and populated it with 32 selected items rated the highest (> 4.0 points with standard deviation < 1.0) in the online study. The resulting questionnaire[15] has six dimensions measuring (1) overall reaction, (2) perceived effectiveness, (3) system capabilities, (4) learnability, (5) visuals/displays and animacy, (6) real-time feedback. The questionnaire allows to evaluate a system's functionality related to multimodality (items in dimension 3 and 5) and tutoring capabilities (dimension 6). The questionnaire is used to perform user-based evaluation and is evaluated on internal consistency reliability (see next Section 6.2.2).

6.2.2 User-based evaluation: perception vs performance

The Virtual Negotiation and Virtual Debate Coaching systems were evaluated measuring usability in terms of effectiveness, efficiency and satisfaction. Previous research suggests that there are differences in perceived and actual performance [Nielsen, 2012]. Performance and perception scores are correlated, but they are different usability metrics and both need to be considered when conducting quantitative usability studies. In our design, subjective perception of effectiveness, efficiency and satisfaction were correlated with various performance metrics and interaction parameters to assess their impact on the qualitative usability properties. We computed bi-variate correlations to determine possible factors impacting user perception of system usability and the derived performance metrics and interaction parameters from logged and annotated evaluation sessions.

The perceptive assessments come from the user satisfaction judgments on different aspects after interacting with the system. The questionnaire designed for this purpose is, first, evaluated on internally consistency and reliability measuring Cronbach's alpha. The internal consistency of the factors (dimensions) were: (1) overall reaction, alpha=0.71; (2) perceived effectiveness, alpha=0.74; (3) system capabilities, alpha=0.73; (4) learnability, alpha=0.72; (5) visuals and animacy, alpha=0.75; and (6) real-time feedback, alpha=0.82. All alpha values were > 0.7, so we can conclude that all factors have sufficient internal consistency reliability.

As part of the user-based evaluation, users were asked to provide an overall rating of the system that they interacted with, using six bipolar negative-positive adjective pairs such as frustrating-satisfying, difficult-easy, inefficient-efficient, unnatural-natural, rigid-flexible and useless-useful, rated on 5-points Likert scales. Correlations between the mean overall satisfaction (3.64) and each of the other factors was measured as follows: effectiveness, $r = .79$; system capabilities, $r = .59$; learnability, $r = .87$; visuals and animacy, $r = .76$; and feedback, $r = .48$. Thus, users appreciate when the system effectively meets their goals and expectations and supports them in completing their tasks, is easy to learn how to interact with and offers flexible input and output processing and generation in multiple modalities.

As performance metrics, system and user performance related to task completion rate[16]

[15] The system demo video and usability questionnaire is available online at `https://www.lsv.uni-saarland.de/index.php?id=72`

[16] We consider the overall negotiation task as completed if parties agreed on all four issues or parties came to the conclusion that it is impossible to reach any agreement.

and its quality[17] were computed. We also compared system negotiation performance with human performance on the number of agreements reached, the ability to find Pareto optimal outcomes, the degree of cooperativeness, and the number of negative outcomes[18]. It was found that participants reached a lower number of agreements when negotiating with the system than when negotiating with each other, 66% vs 78%. Participants made a similar number of Pareto optimal agreements (about 60%). Human participants show a higher level of cooperativity when interacting with the system, i.e. 51% of the actions are perceived as cooperative. This may mean that humans were more competitive when interacting with each other. A lower number of negative deals was observed for human-agent pairs, 21% vs 16%. Users perceived their interaction with the system as effective when they managed to complete their tasks successfully reaching Pareto optimal agreements by performing cooperative actions but avoiding excessive concessions. No significant differences in this respect were observed between human-human and human-system interactions.

As for efficiency, we assessed temporal and duration dialogue parameters, e.g. time elapsed and number of system and/or user turns to complete the task (or sub-task) and the interaction as a whole. We also measured the system response time, the silence duration after the user completed his utterance and before the system responded. Weak negative correlation effects have been found between user perceived efficiency and system response delay, meaning users generally found the system reaction and the interaction pace too slow. Dialogue quality is often assessed measuring word and sentence error rates [Walker et al., 1997, López-Cózar et al., 2006] and turn correction ratio [Danieli and Gerbino, 1995]. Many designers, however, noticed that it is not so much how many errors the system makes that contributes to its quality, but rather the system's ability to recognise errors and recover from them. This contributes to the perceived system robustness and is appreciated by the users. Users value if they can easily identify and recover from their own mistakes. All system's processing results were visualised to the user in a separate window, which contributes to the system observability. System's and user's applied repair and recovery strategies are evaluated by two expert annotators and agreement was measured in terms of kappa. Repairs were estimated as the number of corrected segments, recoveries as the number of regained utterances which were partially failed at recognition and understanding, see also [Danieli and Gerbino, 1995]. While most annotators agreed that repair strategies were applied adequately, longer dialogue sessions due to frequent clarifications seem to be undesirable.

6.2.3 Evaluating the Virtual Negotiation Coach

In our evaluation experiment, 28 participants aged 25-45, professional politicians or governmental workers, were interacting with the VNC system for an hour. Nine negotiation scenarios were used, based on different negotiators preference profiles. Users ('trainees')

[17] Overall task quality was computed in terms of number of *reward points* the trainee gets at the end of each negotiation round and summing up over multiple repeated rounds; and *Pareto optimality* (see footnote 5).

[18] We considered negative deals as flawed negotiation action, i.e. the sum of all reached agreements resulted in an overall negative value, meaning that the trainee made too many concessions and selected mostly dispreferred bright 'orange' options (see Figure 4.9).

were assigned a City Councilor role and a random scenario. All sessions were recorded and numerous interaction parameters logged.

The VNC is evaluated to be relatively easy to interact with (4.2 Likert points). However, users found an instruction round with a human tutor prior to the interaction useful. Most users were confident enough to interact with the system of their own, some of them however found the system too complex and experienced difficulties in understanding certain concepts/actions. A performance metric which was found to negatively correlate with system learnability is user response delay, the silence duration after the system completed its utterance and the user proposed a relevant dialogue continuation. Nevertheless, the vast majority of users learned how to interact with the system and complete their tasks successfully in the consecutive rounds. We observed a steady decline in user response delays from round to round.[19]

Users appreciated the system's flexibility. The system offered the option to select continuation task actions using a graphical interface on a tablet in case the system processing failed entirely. The use of concurrent multiple modalities was positively evaluated by the users. It was always possible for users to take initiative in starting, continuing and wrapping up the interaction, or leave these decisions to the system. At each point of interaction, both the user and the system were able to re-negotiate any previously made agreement.[20]

As overall satisfaction, the interaction was judged to be satisfying, rather reliable and useful, however, less natural (2.76 Likert points). The latter is largely attributed to rather tedious multimodal generation and poor avatar performance. System actions were judged by expert annotators as appropriate[21], correct[22] and easy to interpret. Other module-specific parameters reflecting widely used metrics computed by comparing system performance with reference annotations were various types of error rates, accuracy, and κ scores measuring agreement between the system performance and human annotations of the evaluation sessions. Recognition and interpretation mistakes turned out to have moderate negative effects on the user satisfaction. Table 4.9 summarises the results.

Satisfaction questionnaires were constructed in such a way that, along with overall user satisfaction, we could also evaluate the system's tutoring performance. Participants indicated that system feedback was valuable and supportive. However, they expected more visual real-time feedback and more explicit summative feedback on their learning progress. Most respondents think that the system presents an interesting training skills application and would use it as a part of their training routine.

[19] For now, this is only a general observation. The metric will be taken into consideration in future test-retest experiments.

[20] Performance metrics related to initiative and task substituitivity aspects and their impact on the perceived usability will be an issue for the future research.

[21] System action is appropriate given the context if it introduces or continues a repair strategy.

[22] System action is considered as correct if it addresses the user's actions as intended and expected. These actions exclude recovery actions and error handling.

Usability metric	Perception	Performance		R
	Assessment	Metric/parameter	Value	
effectiveness (task completeness)	mean rating score effectiveness 4.08	Task completion rate[16]; in %	66.0	.86*
		Reward points[17]; mean, max.10	5.2	.19
effectiveness (task quality)		User's Action Error Rate (UAER, in %)[18]	16.0	.27*
		Pareto optimality[17]; mean, between 0 and 1	0.86	.28*
		Cooperativeness rate; mean, in %	51.0	.39*
efficiency (overall)	mean rating score efficiency 4.28	System Response Delay (SRD); mean, in ms	243	-.16
		Interaction pace; utterance/min	9.98	-.18
		Dialogue duration; in min	9:37	-.21
		Dialogue duration average, in number of turns	56.2	-.35*
efficiency (learnability)	3.3 (mean)	User Response Delay (URD); mean, in ms	267	-.34*
efficiency (robustness)	3.2 (mean)	System Recovery Strategies (SRS) correctly activated (Cohen's κ)	0.89	.48*
		User Recovery Strategies (URS) correctly recognised (Cohen's κ)	0.87	.45*
efficiency (flexibility)	3.8 (mean)	Proportion spoken/on-screen actions mean, in % per dialogue	4.3	.67*
satisfaction (overall)	aggregated per user ranging between 40 and 78	ASR Word Error rate; WER, in %	22.5	-.29*
		Negotiation moves recognition accuracy, in %	65.3	.39*
		Dialogue Act Recognition; accuracy, in %	87.8	.44*
		Correct responses (CR)[22] relative frequency, in %	57.6	.43*
		Appropriate responses (AR)[21] relative frequency, in %	42.4	.29*

Table 6.2: Summary of evaluation metrics and obtained results in terms of correlations between subjective perceived system properties and actions, and objective performance metrics (R stands for Pearson coefficient; * = statistically significant ($p < .05$)

Usability metric	Example of trainee's survey entry	M	SD
effectiveness (task success)	I completed my task successfully	4.95	0.72
effectiveness (task quality)	I achieved all my goals	3.35	0.92
efficiency	The system feedback was mostly timely	3.4	1.05
	System feedback was valuable	3.7	0.91
	System feedback made me more aware of my performance	3.45	1.2
	System provided enough feedback	3.07	1.4
satisfaction, (QUIS) [Chin et al., 1988]	I found the interaction with the system natural	3.95	1.15
	I found the interaction with the system engaging	4.7	0.75
	I found the interaction with the system useful	3.95	0.84
	I would use the system in my training routine	4.37	0.86

Table 6.3: Results evaluating effectiveness, efficiency and user satisfaction. M = Mean; SD = Standard Deviation.

6.2.4 Evaluating the Virtual Debate Coach

We performed trainee-based evaluation experiments involving 40 trainees (male and female aged between 14 and 20 years). We did not aim at the trainee's learning gain assessment, which has been performed in a separate study involving the complete 'learner journey' scenario, see also [Van Helvert et al., 2016] and [Koryzis et al., 2016]. The main evaluation goal of the presented study was to assess system performance in the trainee-based setting, including assessing types, granularity, amount and timing of coaching interventions expected to lead to the best learning outcome. For this purpose, participants debated in pairs as described in Section 4.4. A debriefing stage included filling in questionnaires and discussion rounds with trainees and tutors. Questionnaires were constructed in such a way that, along with overall trainee satisfaction, we could also link their judgments to the system's coaching interventions. Trainee judgments were presented in a 1-5 Likert scale. Each session lasted 60-90 minutes including preparation, interaction and filling in a questionnaire. The discussion round involved all participants and tutors after all sessions were completed.

The VDC generated real-time 'in-action' feedback on presentational and interactive aspects such as speech volume, speaking rate, hand and arm position, posture shifts, and turn taking and time management behaviour, i.e. interruptions, overlapping speech and arguments longer than >1 minute were discouraged. Full session recordings, system recognition and processing results, as well as the generated 'in-action' feedback were logged and converted to `.anvil` format for using Anvil tool to view, browse, search, replay and edit debate sessions. This allows automatic generation of the VDC summative feedback to be discussed in 'about-action' training sessions. Moreover, the implemented prediction models can be edited by debaters and tutors on the fly, and corrected annotations can be used to retrain the system.

Table 6.3 summarises the results and shows consistently positive participant feedback for almost all the questions, however, with different deviations from the mean. High task completion rate along with positive effect on skill training is reported. Trainees indicated however that system feedback was sometimes hard to interpret. Most participants found that the system generated too much feedback; such a large amount was difficult to process and distracted from the debate interaction. Trainees also expected more real-time feedback and summative feedback on learning progress.

6.3 Summary

The Virtual Debate and Negotiation Coaches were designed and implemented on the basis of the theoretical frameworks and empirical findings reported in this work. System capabilities are realised through specific modules and include multimodal and interpretation including fusion, dialogue management with the integrated cognitive task agents, and multimodal communicative behaviour generation. Possible extensions are foreseen at the level of the application domain, use case or technical solutions including novel future devices, sensors and processing hardware and algorithms. Following the formal requirements collected within the project, the integration platform was developed, which supports distributed system integration and asynchronous messages-processing inter-module communication using the ZeroMQ library.

To assess the overall system performance, in particular to assess the value of the integration of a cognitive task agent into a dialogue manager, user-based system evaluation was carried out in a series of experiments. For this purpose, an approach to multimodal dialogue system evaluation was proposed which is compliant with the available ISO standards on usability and qualitative metrics for effectiveness, efficiency and satisfaction. A prototype questionnaire containing 110 items was designed, based on established measures and best practices for the usability evaluation of interactive systems and interfaces. Potential questionnaire items were rated by respondents. Eight factors were selected as having a major impact on the perceived usability of a multimodal dialogue system and related to task success, task quality, robustness, learnability, flexibility, likeability, ease of use and usefulness (value). As a result, an internally consistent and reliable questionnaire with 32 items (Cronbach's alpha of 0.87) was extracted. This questionnaire was used to evaluate the Virtual Debate and Negotiation Coaching systems.

Performance metrics and interaction parameters were either automatically derived from logfiles or computed using reference annotations. Perception and performance were correlated to be able to quantify usability.

In Negotiation and Debate scenarios, it was observed that the overall system usability is mostly determined by the user satisfaction with the task quality, by the robustness and flexibility of the interaction, and by the quality of system responses.

Additionally, to assess natural multimodal argumentative behaviour, a set of criteria was defined that helps to explain observed regularities and induced rules, strategies and constraints for the generation, assessment and correction of trainees' debate performance. Experiments of various type supported fairly reliable identification of multimodal markers, and their linking to assessments of argument structure, quality and delivery aspects.

The ambitious vision of the VDC and VNC presents a significant number of challenges. A fully automatic system that is able to understand complex negotiation strategies and natural debate arguments accurately enough to achieve human-like performance has not been achieved yet due to certain limitations in sensor tracking, speech recognition, and natural language processing technologies. Also since a data-oriented approach for modelling of many dialogue phenomena has been deployed, the current quantity and quality of multimodal data was insufficient for training statistical machine learning algorithms.

There is a lot of room for further research. Our main goal is to advance in achieving *immersive* coaching, when the system will enter, exit and re-enter different modes, e.g. monitoring, mirroring, exercising, reflecting, guiding and freestyle modes, see next chapter for more elaborate discussion.

Chapter 7

Conclusions and perspectives

In this chapter we formulate the main conclusions of the research reported in this thesis, and indicate perspectives and directions for future research that can build on this work.

7.1 Conclusions

Cognitive modelling of metacognition processes in dialogue This thesis starts from the consideration that dialogue is a complex activity. A conversational interactant mostly operates in real time, trying to interpret intentions of the speaker as the dialogue is produced. In dialogue both partners cooperate in the co-construction of the meaning. Conversing is different, in this respect, from reading where a reader can browse a text at leisure, scan across the lines and integrate the words to build a personal representation of the meaning as set down by the author. Dialogue participants have many, often parallel tasks to perform during interaction. While performing task(-s), e.g. exchanging certain information, instructing another participant, negotiating an agreement, expressing opinions and defending them, dialogue participants need to connect and organise ideas, fill gaps in their knowledge, evaluate evidence, argue with new information, test and modify, predict, clarify, generate questions, learn new concepts, make unexpected connections, reflect, analyse, synthesise and loop back. Among these things, dialogue participants have constantly to 'evaluate whether and how they can (and/or wish to) continue, perceive, understand and react to each others intentions' [Allwood, 2000]. They monitor contact and attention, elicit feedback, share information about the processing of each others messages.

For the dialogue system to be able to act as a plausible dialogue partner and to show human-like interactive behaviour, it needs to *learn*. Learning involves strengthening of existing knowledge, compilation of new rules, collection of episodic experiences to improve future decisions. Learning also requires assessing why a particular solution worked or not, and manipulation of the task representation accordingly. This is known as self-regulated learning (SRL) which involves metacognitive skills development. People learn to understand, control and manipulate their own cognitive processes: monitor the degree to which they understand new information, recognise failures to comprehend, employ effective help-

seeking and repair strategies, adjust their learning process to feedback from others, and maintain the attitudes necessary to invoke and employ learning strategies. Common among these processes is that they lead to gradual improvement of performance over time.

One of the key functions of metacognition is to improve learning. Since metacognition, despite what it's name may suggest, is considered as a cognitive process at the level and along other cognitive processes [Salvucci and Taatgen, 2008], it is something that can be learned, just as any other skill. Therefore, a good way to provide a dialogue system with metacognitive abilities is to understand how people acquire such skills while training them. We focused on three types of metacognitive skills: *monitoring*, *reflection* and *regulation*. However, longer term goal may be that of predicting dialogue participants' knowledge and intentions and for the system to show dialogue behaviour that can be considered proactive. Proactive behaviour is achieved through anticipating future demands of tasks in order to improve overall task performance and optimise sub-task switching. Such predictive processing is considered as central to many cognitive functions, and seems to be required in the processing of information from one's environment. Proactive behaviour can be trained and a dialogue system needs to become a proactive learner whose performance improves over time.

As a theoretical basis for developing such an account, the Cognitive Task Analysis (Chipman et al., 2000) and the ACT-R cognitive architecture (Anderson, 2007) were applied. For the basic model of (meta)cognitive processes, observable dialogue behaviour was related to the task performance and goal structures resulting in the Debate Coach Agent performing as an Observer, Mirrorer and Tutor in the debate scenario. ACT-R provides a simulation system for general cognitive processing, and has already been used to model metacognition (e.g. van Rij, van Rijn and Hendriks, 2010). The ACT-R Cognitive Task Agent is equipped with Theory of Mind skills (Premack and Woodruff, 1978) and is able to use its task knowledge not only to determine its own actions, but also to interpret the human partner's actions, and to adjust its behaviour to whom it interacts with. The artificial agents are trained to employ instance-based learning to decide about its own actions and to reflect on the behaviour of the opponent. The Negotiation Task Agent performs as an Experiencer and a Tutor in the negotiation scenario. We have shown that the integrated cognitive task agents perform actions and make decisions comparable to those of human.

Advances in dialogue management: data-driven design In the future, multimodal interactive systems will operate on huge, dynamic, heterogeneous data streams, providing powerful possibilities for adaptive and flexible navigation and visualisation (Renals et al., 2014). Substantial progress has already been made in this field, generating a plethora of data which is often available for research and applications. This boosted the development of data-intensive applications such as Smart Assistants from Apple (Siri), Google (Google Assistant), Amazon (Alexa) Microsoft (Cortana) and Ebay (ShopBot) or chat commerce applications integrated into Facebook Messenger, WhatsApp, Talk, and WeChat. The amount of real user data available to developers of these systems contributed significantly to their success and robustness. Data became essential in advancing the state of the art in many research fields related to human language technologies. Paired with increased computational power, it enabled the development of Artificial Neural networks (ANNs) and Deep

Learning (DL), which had significant impact on Artificial Intelligence (AI) and Language Technologies (LT).

The sheer amount of available data does not necessarily lead to knowledge, but rather provides opportunities to gain more and better insights. Obtaining relevant high quality data is not enough if we want to exploit the full potential of the new digital world to the largest extent possible. Data from various sources needs to be connected in a sensible way to give us deeper insights into the nature of different events, objects or processes. Since the data has to be realistic and appropriate for the respective use case we have to deal with real user data, which is inherently heterogeneous: multimodal, incomplete, inconsistent and often noisy (unreliable). System design processes still require skilled manual labour to embed intelligence into the data-driven dialogue modelling. Skills to analyse and extract the right data and metrics and skills to use that data to inform the decisions to be taken in the system design process. In other words, there are expert and novice users who formulate requirements and evaluate the products, and there are experts (domain experts, system designers or annotators) who analyse and interpret the collected raw data which is often specified as an annotation process - the process of adding linguistic information to the primary data. For the purpose of dialogue modelling, *dialogue act* annotations are performed. Most existing dialogue models (in fact, all except perhaps for end-to-end approaches) make use of semantic information of this type. Numerous efforts have been undertaken to standardise dialogue act annotation models, schemes and evaluation criteria contributing to the creation of interoperable dialogue resources. Standards do not only enable interoperable resource creation and re-use of existing not-interoperable ones, but also allow comparing analysis results from different studies, and validate different approaches.

To pursue this line of development, we proposed the Continuous Dialogue Corpus Creation (D3C) methodology. In this approach, a corpus serves as a shared repository for analysis and modelling of interactive dialogue behaviour, and for implementation, integration and evaluation of dialogue system components. All these activities are supported by the use of ISO standard data models including annotation schemes, encoding formats, tools, and architectures. Standards also facilitate practical work in dialogue system implementation, deployment, evaluation and re-training, and enable automatic generation of adequate system behaviour from the data. The Dialogue Act Markup Language specified in the ISO 24617-2 standard (DiAML) is used as an interface language between modules of a multimodal dialogue system.

From the system development perspective, many learning systems require long training and large set of examples. The D3C methodology in combination with Cognitive Task analysis and IBL-based cognitive task models facilitated the cognitive task agent design based on rather limited real or simulated dialogue data - about 2.5 hours of real dialogue data for each scenario was used. The agents are supplied with the initial list of assessment criteria and/or state-action templates encoding domain expert knowledge and the agent's preferences and strategies, and the agents collect interactive experiences and learn from them.

Advances in dialogue management: effective implementation of complex models The implementation introduced a theoretical novelty in instance-based learning for Theory

of Mind skills and integrated this in the dialogue management of an interactive cognitive tutoring system. Cognitive Agents are integrated as Task Agents into the multi-agent ISU-based dialogue management module. The Dialogue Manager operates on the basis of the multidimensional context model which has a five-component structure in order to represent the participants' multimodal information state as adequate as possible, accounting for multiple factors that influence interpretation and generation of rich multimodal dialogue behaviour. The system shows multi-tasking behaviour and can play different dialogue roles addressing various task-specific and interactive goals, and expectations simultaneously. The approach leads to a knowledge-rich representation of the participants' information states and highly flexible dialogue management strategies. Moreover, it offers possibilities for various future extensions.

The proposed multi-agent Dialogue Manager architecture allows separating task-related and dialogue control actions. This enables the application of sophisticated models along with a flexible architecture in which various (including alternative) modelling approaches can be combined. The approach was illustrated by integrating two different Cognitive Task Agents - a Debate Coach Agent and a Negotiation Task Agent.

Dialogue system evaluation As a part of the design of any interactive application, evaluations need to be performed in order to assess the success of the developed services and systems by matching performance criteria to user expectations, needs and requirements. Evaluation results serve to inform designers about functional and non-functional deficiencies. In addition, a well-designed evaluation may give a good indication whether proposed solutions would be accepted by potential users.

User-based dialogue system evaluation is expensive. Automated metrics, however, do not always provide meaningful comparison, and rather penalise than reward novelty and creativity [Georgila et al., 2018b]. Automated unsupervised metrics, e.g. BLEU, ROUGE, METEOR, do not correlate well with human judgments of dialogue quality (Liu et al., 2016).

When evaluating an interactive application, user judgments provide the most valuable insights how well the designed product meets their expectations and needs. It has been acknowledged that user involvement at all system development stages, in particular in evaluation activities, ensures that the product designed is usable, see e.g. Cooperative Design [Greenbaum and Kyng, 1992], Participatory Design [Spinuzzi, 2005], User-Centered Design [Blackburn and Cudd, 2010].

We proposed to assess multimodal dialogue system performance by relating the relative contribution of various objective parameters and subjective factors to the usability of a dialogue system as defined by the ISO 9241-11 and ISO/IEC 9126-4 standards. To quantify usability, subjective perception of effectiveness, efficiency and satisfaction were correlated with various performance metrics and interactive parameters, e.g. error rates, accuracy and precision, number of (in)appropriate system responses, recovery strategies, and interaction pace. The standard usability approach provides a useful decomposition of the usability concept into several dimensions (factors), enabling a clear mapping of system performance to distinctive usability perception aspects, with the advantage of being able to assess the im-

pact of different items on usability perception instead of simply summing up or averaging to compute an overall satisfaction score.

General conclusions Coming to more general conclusions, the first conclusion is that the application of a multidimensional view on communication in combination with cognitive modelling of the relevant information related to learning, multitasking, prediction and proactive cognitive control leads to a better understanding of human dialogue behaviour and enables better computational modelling of multimodal dialogue. The obtained insights in the nature and types of metacognitive processes and multifunctionality of dialogue contributions incorporated in a dialogue system results in adequate understanding and generation of task-related actions in non-sequential multimodal interactions, it facilitates decision making in a way that is comparable to human behaviour of this type. A multidimensional approach to dialogue modelling opens the perspective for rich human-system interaction. It supports more accurate understanding and better multimodal and multi-tasking behaviour.

A second general conclusion is that the use of fundamental concepts and insights from cognitive science and dialogue theory is generally useful for an adequate analysis of human dialogue behaviour, for modelling this behaviour, and for the design of dialogue systems. In particular, it may be observed that context-driven dialogue understanding and generation makes use of assumptions concerning rational and cooperative behaviour of the dialogue participants, showing that such assumptions are useful, if not indispensable in computational modelling of dialogue.

A third general conclusion is that the analytical and empirical studies reported in this thesis have contributed to the creation of dialogue corpora annotated with interoperable semantic information, many of them included into the DialogBank repository (Bunt et al., 2016) and released to the research community under LDC[1] and ELRA[2] licenses. The detailed investigation of multimodality of dialogue contributions, of the domain-specific semantics of functional segments, and of semantic relations between them, has contributed to certain revisions of the ISO standard 24617-2 for dialogue act annotation proposed recently, see e.g. [Bunt et al., 2017a].

To summarise, the main ideas, concepts and assumptions of a theory of dialogue, such as the 'information-state' theory in general and Dynamic Interpretation Theory in particular, which consider the meaning of communicative behaviour in terms of changes in the participants' state of information upon successful communication, combined with a multidimensional view on dialogue communication and elaborate cognitive modelling of task-related learning and decision taking processes, open the way to design effective, efficient and adaptive dialogue systems that are flexible enough to exploit the full potential of spoken and multimodal interaction.

[1] https://catalog.ldc.upenn.edu/LDC2017S11

[2] http://catalogue-old.elra.info/product_info.php?products_id=1317

7.2 Perspectives

Having produced the results and conclusions discussed in the previous section, this thesis concludes by suggesting new opportunities and directions for future work.

Beyond task: social cognition Knowledge and experience is gathered via learning and interaction, observing and imitating others, understanding and following best practices in everyday situations. Our interactions are more than the exchange of information and offers, decision making or problem-solving; they involve a wide range of aspects related to feelings, emotions, social status, power, and interpersonal relations, and the context. For many real-life interactive situations, it is important to maintain good relations, e.g. build trust over time. In contrast, interactive processes that lead to poor communication, polarisation and conflict escalation may be observed in interactions over social barriers such as interactions involving participants of different genders, races or cultures.

Social cognition constitutes the primary adaptations in our cognition. These adaptations can be seen in several developments. Relating to action is the development of increasingly sophisticated representations of agency and self, together with increasingly powerful abilities for social mirroring, imitation, and cooperative action. Relating to theory of mind is the development of the abilities to establish joint attention and represent the minds of others. Relating to communication is the development of remarkable new abilities to use language, establish social groups, create culture, and archive cultural bodies of knowledge. For all these reasons, *understanding human cognition successfully requires understanding its coupling to the social environment.* Consequently, designers will target the *design of socially embedded systems.*

Immersive experiences It has been shown that digital immersion can enhance learning, user experience, motivation and by this user acceptance, in three ways: by providing (1) multiple perspectives; (2) situated learning; and (3) transfer (Dede, 2009; Lessiter et al., 2001; Sadowski and Stanney, 2002). While the vision for our Virtual Coaches may not be considered "digitally immersive" in the traditional sense, it can be said to incorporate these three types of experience.

An immersive and highly individualised coaching experience can be achieved by effective use and elaborate analysis of interaction data, applying advanced affective signal processing techniques and rich domain knowledge. A comprehensive account of users' feelings, motivations, and engagement in the dialogue model will form the foundation for a new generation of interactive tutoring systems. Motivation and emotions play a key role in interaction in general and coaching contexts in particular. A direction that is not yet explored to the full is to optimise for users' motivation and engagement in a system, as opposed to optimise for pure functional efficiency.

Through a deep understanding of leading motivation theories: Flow Theory (Csikszentmihalyi, 1997), Keller's Motivational Model (Keller, 2009), Self-Determination Theory (Deci and Ryan, 2011) and Social Cognitive Theory (Bandura, 2011), different techniques can be combined to form desired and joyful experiences for learners, to keep them engaged for longer periods of time, building meaningful relationships between them, and developing their creative potential.

Users, interacting with any system for a longer period of time, generate a significant amount of interaction data that can be measured and utilised to create individualised experiences. This interaction data can be used to enrich information encoded in the affective and mental states. Only a system that understands the user can provide individualised support and only a system that is able to react in real time to numerous tracking data is able to provide a truly immersive experience. For example, there is evidence that an immediate reflective response to a puzzling event in the practice plays a crucial role in learning. It seems that on-the-spot, amplified and explicit reflection, and real-time feedback bears on a continuous awareness and appraisal of the act of learning itself more than on problem-solving flash-light insights.

A future ambition would be to make significant advances by *combining a strong theory-based framework for motivation and engagement with advanced monitoring, beyond-state-of the art affective signal processing, and access to real-users interactive data of sufficient amount and quality.*

Multimodality Closely related to the previous point, a fundamental idea of future developments is to observe and analyse dialogue acts across all available modalities rather than focusing on a single or limited number of modes. This analysis should help to make more holistic propositions on the actual performance and behaviour of learners and thus provide better instructional feedback on how to adapt individual strategies. In future versions of the system the principles discussed here will be transferred to other multimodal aspects captured by the system, e.g. emotional information, sentiment analysis, etc.

The digital and physical worlds are currently merging, opening new possibilities for us to interact with our environment, as well as for our environment to interact with us. Everyday objects and smart toys, which previously did not seem aware of the environment at all, are turning into smart devices with sensing, tracking or alerting capabilities. Approaches such as the "beyond desktop" metaphor [Kaptelinin and Czerwinski, 2007] for design of interfaces, and "disappearing" or ubiquitous computing [Streitz, 2001] for design of digital technology have existed for over two decades and can be profitably explored for our application domain - metacognitive skills training.

Dialogue management

Shared control There are several ways of implementing adaptive strategies into human-machine interaction. The most frequently used one is to control the sequence, content etc. of all interactional activities by the system (=program controlled adaptation). A system's proactive control has been modelled in this work. A second way of adaptation is based on the idea that the user has to take an active part in interaction, and thus is provided with a choice of interactional activities (=user controlled adaptation). Both ways have been found to have their limits: If the system takes the lead in adaptation, by implication the user can only influence task performance, e.g. type of information requested. If the user is responsible for the adaptation, they need to have sufficient background knowledge to decide which information, resources and activities would be best for them. Unfortunately, users often lack these skills. We propose a third way of adaptation, which combines both program and user control (= shared control adaptation). Shared-control adaptive systems

have been recommended as a promising way of promoting the acquisition of not only task specific domain knowledge and skills but also higher-order self-regulation knowledge, skills and/or strategies (i.e., self-evaluation of learning processes and outcomes).

In the future, we will explore shared control between the user and the cognitive agents, which offers a variety of new possibilities for adaptation and raises new research issues. Building on the models designed in this work, we will move further away from widely used linear learning and interaction sequences and scripting to non-linear interactive scenario building with agents, dialogues and branches leveraging deep interactive learning, and building competences. Agents will not only learn from interactions (as simulated in this work), but they will be taught by humans. Learned behaviours (including behaviour related to social aspects and affected agent's states) will be saved into persistently growing libraries of instances. The user control will be achieved by explicit teaching, where either user or a domain expert can teach the system to behave in certain way. The goal is to design a unique approach of interactive shared control, for a reliable and convenient way to receive information through active interaction combining system and user control adaptation methods.

Multi-perspective dialogue To enable fundamentally deeper understanding of metacognitive processes and the nature of the acquisition of such skills, the system needs shared and varied responsibilities of observing, monitoring, experiencing and executing different tasks, by presenting similar materials in multiple contexts enabling self-reflection, by becoming aware of different strategies and how they work. To a certain extent this has been realised in this work. However, we think there is much more potential here, in particular in the combination with the concept of shared control presented above.

The next system generation will have an ability to lead/be engaged in multimodal social conversation with a multi-perspective support for an immersive coaching experience. An advanced multi-perspective approach will allow both, the system and the user, to switch dynamically and in real-time between different character roles and dialogue/task responsibilities, cease the interaction, resume it re-entering the same mode or replay it in a different role (from different perspective) and/or mimicking the partner's behaviour.

Incremental parallel processing There is overwhelming psycholinguistic evidence that human language processing is incremental. Humans construct syntactic, semantic, and pragmatic hypotheses on the fly, while receiving written or spoken input. If a language understanding system is also able to interpret user utterances incrementally, the system will be able to show interactive behaviour that is natural to its users. For example, using dialogue phenomena such as backchannelling (providing feedback while someone else is speaking, also called active listening), providing the completion of a user utterance that he is struggling to finish, and even interrupting the user, for example, to correct him, urgently express an alternative opinion, disagreement or request for clarification, or, by contrast, to express appreciation or encouragement.

In dialogue system design, enabling incremental interpretation will allow the system to respond more quickly, by minimising the delay between the time the user finishes and the time the utterance is interpreted.

Recently, systems are developed where any minimal input triggers the system's processing which continues increment-by-increment till the complete input is recognised (Schlangen and Skantze, 2009). This creates possibilities for the system to show more interactive and pro-active behaviour (e.g. backchannelling, interrupting and completing the partner) and to minimise system response time, see [Aist et al., 2007].

Full incremental interpretation and understanding of dialogue utterances, in our view, can be achieved when the tasks, such as automatic speech recognition, segmentation, lexical search, syntactic and semantic parsing, pragmatic interpretation, can be realised incrementally and in close interaction with each other. There is no agreement however on the nature of its minimal units, i.e. increments. There is also no evidence that for all processing steps/types, increments should be of the same nature and size. For example, for dialogue act classification, which has a higher level of semantic abstraction, a token/word-based approach might be not the most adequate one. Bigger units may form the basis for incremental dialogue act processing. Such units, chunks, can be motivated by prosody, syntax or semantics. The preliminary results show that compared to token/word-based incremental classification a syntactic and semantic chunk-based classification produces better results on manual transcriptions (negotiation scenario), see [Ebhotemhen et al., 2017].

Our vision: given a context model that is monitored and updated during the dialogue, an incremental Dialogue Manager starts generating dialogue acts in several dimensions simultaneously even before the user finishes his turn. Following the incremental approach participants' information-states will be updated based on available partial input interpretation. These updates will be kept in the pending context and evaluated for consistency. If inconsistencies occur, this may also mean that an initial interpretation is wrong and another hypothesis may be considered.

Bibliography

[Ahn, 2001] Ahn, R. (2001). *Agents, Objects and Events. A computational approach to knowledge, observation, and communication. PhD Thesis*. Eindhoven University of Technology, The Netherlands.

[Aist et al., 2007] Aist, G., Allen, J., Campana, E., Gomez Gallo, C., Stoness, S., Swift, M., and Tanenhaus, M. K. (2007). Incremental understanding in human-computer dialogue and experimental evidence for advantages over nonincremental methods. In Arstein, R. and Vieu, L., editors, *Proceedings of the 11th Workshop on the Semantics and Pragmatics of Dialogue*, pages 149–154, Trento, Italy.

[Aleven et al., 2006] Aleven, V., Mclaren, B., Roll, I., and Koedinger, K. (2006). Toward meta-cognitive tutoring: A model of help-seeking with a cognitive tutor. *International Journal of Artificial Intelligence in Education*, 16:101–128.

[Allen, 1983] Allen, J. (1983). Recognising intentions from natural language utterances. In *Computational Models of Discourse, Brady, M., and Berwick, R.C. (eds)*, pages 107–166. MIT Press, Cambridge, MA.

[Allen and Core, 1997] Allen, J. and Core, M. (1997). Draft of DAMSL: Dialog Act Markup in Several Layers. Available at `http://www.cs.rochester.edu/research/cisd/resources/damsl/`.

[Allen et al., 2001] Allen, J., Ferguson, G., and Stent, A. (2001). An architecture for more realistic conversational systems. In *Proceedings of the 6th international conference on Intelligent user interfaces*, pages 1–8. ACM.

[Allen and Perrault, 1980] Allen, J. and Perrault, C. (1980). Analyzing intention in utterances. *Artificial intelligence*, 15(3):143–178.

[Allen et al., 1995] Allen, J., Schubert, L., Ferguson, G., Heeman, P., Hwang, C., Kato, T., Light, M., Martin, N., Miller, B., Poesio, M., and Traum, D. (1995). The TRAINS project: A case study in building a conversational planning agent. *Journal of Experimental & Theoretical Artificial Intelligence*, 7(1):7–48.

[Allwood, 1977] Allwood, J. (1977). A critical look at speech act theory. *Logic, Pragmatics and Grammar, Dahl (eds.)*, pages 53–69.

[Allwood, 1992] Allwood, J. (1992). On dialogue cohesion. *Gothenburg Papers In Theoretical Linguistics*, 65.

[Allwood, 1994] Allwood, J. (1994). Obligations and options in dialogue. *THINK Quarterly 3(1)*, pages 9–18.

[Allwood, 2000] Allwood, J. (2000). An activity-based approach to pragmatics. *Abduction, Belief and Context in Dialogue*, pages 47–81.

[Allwood et al., 1992] Allwood, J., Nivre, J., and Ahlsén, E. (1992). On the semantics and pragmatics of linguistic feedback. *Journal of semantics*, 9(1):1–26.

[Allwood et al., 2000] Allwood, J., Traum, D., and Jokinen, K. (2000). Cooperation, dialogue and ethics. *International Journal of Human Computer Studies*, pages 871–914.

[Altmann and Gray, 2008] Altmann, E. and Gray, W. (2008). An integrated model of cognitive control in task switching. *Psychological review*, 115(3):602.

[Amanova et al., 2016] Amanova, D., Petukhova, V., and Klakow, D. (2016). Creating annotated dialogue resources: Cross-domain dialogue act classification. In *Proceedings 9th International Conference on Language Resources and Evaluation (LREC 2016)*. ELRA, Paris.

[Ambady and Rosenthal, 1992] Ambady, N. and Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis.

[Anderson et al., 1991] Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H., and Weinert, R. (1991). The HCRC Map Task corpus. *Language and speech*, 34(4):351–366.

[Anderson, 2007] Anderson, J. (2007). *How Can the Human Mind Occur in the Physical Universe?* New York, NY: Oxford University Press.

[Anderson et al., 2004] Anderson, J., Bothell, D., Byrne, M., Douglass, S., Lebiere, C., and Qin, Y. (2004). An integrated theory of the mind. *Psychological review*, 111(4):1036.

[Anderson and Schooler, 1991] Anderson, J. and Schooler, L. (1991). Reflections of the environment in memory. *Psychological science*, 2(6):396–408.

[Anderson, 2014] Anderson, J. R. (2014). *Rules of the mind*. Psychology Press.

[Anderson and Betz, 2001] Anderson, J. R. and Betz, J. (2001). A hybrid model of categorization. *Psychonomic Bulletin and Review*, 8(4):629–647.

[Andry et al., 1990] Andry, F., Bilange, E., Charpentier, F., Choukri, K., Ponamalè, M., and Soudoplatoff, S. (1990). Computerised simulation tools for the design of an oral dialogue system. Selected Publications, 1988-1990, SUNDIAL Project (Esprit P2218). Commission of the European Communities.

[Annett, 2003] Annett, J. (2003). Hierarchical task analysis. *Handbook of cognitive task design*, 2:17–35.

[Annett et al., 1971] Annett, J., Duncan, K., Stammers, R., and Gray, M. (1971). Task analysis. *HMSO, London. Artman, H.(2000). Team situation assessment and information distribution. Ergonomics*, 43(8):1076–95.

[Aquilar and Galluccio, 2007] Aquilar, F. and Galluccio, M. (2007). *Psychological processes in international negotiations: Theoretical and practical perspectives*. Springer Science & Business Media.

[Arslan et al., 2017] Arslan, B., Taatgen, N. A., and Verbrugge, R. (2017). Five-year-olds systematic errors in second-order false belief tasks are due to first-order theory of mind strategy selection: A computational modeling study. *Frontiers in Psychology*, 8.

[Ashley et al., 2007] Ashley, K., Pinkwart, N., Lynch, C., and Aleven, V. (2007). Learning by diagramming supreme court oral arguments. In *Proceedings of the 11th International Conference on Artificial Intelligence and Law*, ICAIL '07, pages 271–275, Stanford, California. ACM.

[Attardo, 1997] Attardo, S. (1997). Locutionary and perlocutionary cooperation: The perlocutionary cooperative principle. *Journal of Pragmatics*, 27(6):753–779.

[Aust et al., 1994] Aust, H., Oerder, M., Seide, F., and Steinbiss, V. (1994). Experience with the Philips automatic train timetable information system. In *Interactive Voice Technology for Telecommunications Applications, 1994., Second IEEE Workshop on*, pages 67–72. IEEE.

[Austin, 1962] Austin, J. (1962). *How to do things with words*. University Press.

[Ayer, 1966] Ayer, A. J. (1966). *Logical positivism*. Simon and Schuster.

[Azevedo et al., 2002] Azevedo, R., Witherspoon, A., Chauncey, A., Burkett, C., and Fike, A. (2002). MetaTutor: A metacognitive tool for enhancing self-regulated learning. In Pirrone, R., Azevedo, R., and Biswas, G., editors, *Cognitive and Metacognitive Educational Systems: Papers from the AAAI Fall Symposium (FS-09-02)*.

[Azevedo et al., 2009] Azevedo, R., Witherspoon, A., Graesser, A., McNamara, D., Chauncey, A., Siler, E., Cai, Z., Rus, V., and Lintean, M. (2009). MetaTutor: Analyzing self-regulated learning in a tutoring system for biology. In *AIED*, pages 635–637.

[Baker et al., 2006] Baker, R., Corbett, A., Koedinger, K., and Roll, I. (2006). Generalizing detection of gaming the system across a tutoring curriculum. In *Intelligent Tutoring Systems: 8th International Conference, ITS 2006, Jhongli, Taiwan, June 26-30, 2006*, volume 4053 of *Lecture Notes in Computer Science*, pages 402–411. Springer.

[Bandura, 1991] Bandura, A. (1991). Self-regulation of motivation through anticipatory and self-reactive mechanisms. In *Perspectives on motivation: Nebraska symposium on motivation*, volume 38, pages 69–164.

[Bandura, 2001a] Bandura, A. (2001a). Social cognitive theory: An agentic perspective. *Annual review of psychology*, 52(1):1–26.

[Bandura, 2001b] Bandura, A. (2001b). Social cognitive theory of mass communication. *Media psychology*, 3(3):265–299.

[Bandura, 2011] Bandura, A. (2011). Social cognitive theory. *Handbook of social psychological theories*, 2012:349–373.

[Barrett et al., 2007] Barrett, L., Mesquita, B., Ochsner, K., and Gross, J. (2007). The experience of emotion. *Annual Review of Psychology*, 58:373–403.

[Barsalou et al., 2003] Barsalou, L. W., Niedenthal, P. M., Barbey, A. K., and Ruppert, J. A. (2003). Social embodiment. *Psychology of learning and motivation*, 43:43–92.

[Bartneck et al., 2009] Bartneck, C., Kulić, D., Croft, E., and Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International journal of social robotics*, 1(1):71–81.

[Bayer et al., 2017] Bayer, A., Stepanov, E., and Riccardi, G. (2017). Towards end-to-end spoken dialogue systems with turn embeddings. In *Proceedings of the 18th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2516–2520, Stockholm, Sweden. International Speech Communication Association (ISCA), Baixas, France.

[Beach and Connolly, 2005] Beach, L. and Connolly, T. (2005). *The psychology of decision making: People in organizations*. Sage.

[Beard, 2002] Beard, A. (2002). *The language of politics*. Routledge, London.

[Bennacef et al., 1996] Bennacef, S., Devillers, L., Rosset, S., and Lamel, L. (1996). Dialog in the RAILTEL telephone-based system. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, volume 1, pages 550–553. IEEE.

[Bennett and Rudnicky, 2002] Bennett, C. and Rudnicky, A. (2002). The Carnegie Mellon Communicator corpus. In *7th International Conference on Spoken Language Processing, INTERSPEECH 2002*, Denver, Colorado.

[Bereiter, 2005] Bereiter, C. (2005). *Education and mind in the knowledge age*. Routledge.

[Besser, 2006] Besser, J. (2006). A corpus-based approach to the classifiation and correction of disfluencies in spontaneous speech. Master Thesis, Saarland University, Saarland, Germany.

[Bilange, 1991] Bilange, E. (1991). A task independent oral dialogue model. In *Proceedings of the Fifth Conference of the European Chapter of the Association for Computational Linguistics*, pages 83–88, Berlin, Germany. Association for Computational Linguistics.

[Blackburn and Cudd, 2010] Blackburn, S. and Cudd, P. (2010). An overview of user requirements specification in ict product design. *Licences, copyright, patents and business: Making the most of exploitable assets through open*, page 9.

[Boersma and Weenink, 2009] Boersma, P. and Weenink, D. (2009). Praat: doing phonetics by computer. computer program. Available at `http://www.praat.org/`.

[Bohus and Rudnicky, 2003] Bohus, D. and Rudnicky, A. (2003). Ravenclaw: Dialog management using hierarchical task decomposition and an expectation agenda. In *Eurospeech-2003*, Geneva, Switzerland.

[Borst and Anderson, 2015] Borst, J. and Anderson, J. (2015). Using the ACT-R cognitive architecture in combination with fMRI data. In *An introduction to model-based cognitive neuroscience*, pages 339–352. Springer.

[Bos et al., 2003] Bos, J., Klein, E., Lemon, O., and Oka, T. (2003). DIPPER: description and formalisation of an information-state update dialogue system architecture. In *Proceedings of the 4th SIGdial Workshop on Discourse and Dialogue*, pages 115–124.

[Bos and Oka, 2002] Bos, J. and Oka, T. (2002). An inference-based approach to dialogue system design. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–7. Association for Computational Linguistics.

[Boser et al., 1992] Boser, B., Guyon, I., and Vapnik, V. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152. ACM.

[Bothell, 2004] Bothell, D. (2004). ACT-R 6.0 Reference Manual, Working Draft.

[Bovair et al., 1990] Bovair, S., Kieras, D. E., and Polson, P. G. (1990). The acquisition and performance of text-editing skill: A cognitive complexity analysis. *Human-Computer Interaction*, 5(1):1–48.

[Braga and Marques, 2004] Braga, D. and Marques, M. (2004). The pragmatics of prosodic features in the political debate. In *Speech Prosody 2004, International Conference*, pages 321–324, Nara, Japan. ISCA Special Interest Group on Speech Prosody.

[Bridge, 2002] Bridge, D. G. (2002). Towards conversational recommender systems: A dialogue grammar approach. In *Proceedings of the 6th European Conference on Case-Based Reasoning (ECCBR) Workshops*, pages 9–22, Aberdeen, Scotland, UK.

[Brooke et al., 1996] Brooke, J. et al. (1996). SUS-A quick and dirty usability scale. *Usability evaluation in industry*, 189(194):4–7.

[Bruseberg and Shepherd, 2017] Bruseberg, A. and Shepherd, A. (2017). Job design in integrated mail processing. *Engineering Psychology and Cognitive Ergonomics: Volume Five-Aerospace and Transportation Systems*.

[Bunt and Conati, 2003] Bunt, A. and Conati, C. (2003). Probabilistic student modelling to improve exploratory behaviour. *User Modeling and User-Adapted Interaction*, 13(3):269–309.

[Bunt, 1989] Bunt, H. (1989). Information dialogues as communicative action in relation to partner modelling and information processing. In Taylor, M., Neel, F., and Bouwhuis, D., editors, *The Structure of Multimodal Dialogue*, volume 1, pages 47–73. Elsevier, North Holland, The Netherlands.

[Bunt, 1994] Bunt, H. (1994). Context and dialogue control. *THINK Quarterly 3(1)*, pages 19–31.

[Bunt, 1996] Bunt, H. (1996). Interaction management functions and context representation requirements. In Luperfoy, S., Nijholt, A., and Veldhuizen van Zanten, G., editors, *Dialogue Management in Natural Language Systems*, pages 187–198.

[Bunt, 1999] Bunt, H. (1999). Dynamic interpretation and dialogue theory. In Taylor, M., Neel, F., and D., B., editors, *The structure of multimodal dialogue II*, pages 139–166. John Benjamins, Amsterdam.

[Bunt, 2000] Bunt, H. (2000). Dialogue pragmatics and context specification. In Bunt, H. and Black, W., editors, *Abduction, Belief and Context in Dialogue; studies in computational pragmatics*, pages 81–105. John Benjamins, Amsterdam.

[Bunt, 2007] Bunt, H. (2007). Multifunctionality and multidimensional dialogue act annotation. In *Communication - Action - Meaning, A Festschrift to Jens Allwood. E. Ahlsèn et al. (ed.)*, pages 237–259. Göteburg University Press.

[Bunt, 2009] Bunt, H. (2009). The DIT++ taxonomy for functional dialogue markup. In Heylen, H., Pelachaud, C., Catizone, R., and Traum, D., editors, *Proceedings of the AAMAS 2009 Workshop 'Towards a Standard Markup Language for Embodied Dialogue Acts' (EDAML 2009)*, pages 13–25, Budapest.

[Bunt, 2011] Bunt, H. (2011). Multifunctionality in dialogue. *Computer, Speech and Language*, 25:222–245.

[Bunt, 2012] Bunt, H. (2012). The semantics of feedback. In *16th Workshop on the Semantics and Pragmatics of Dialogue (SEMDIAL 2012)*, pages 118–127.

[Bunt, 2014a] Bunt, H. (2014a). Annotations that effectively contribute to semantic interpretation. In *Computing Meaning*, volume 4. Springer, Dordrecht.

[Bunt, 2014b] Bunt, H. (2014b). A context-change semantics for dialogue acts. In Bunt, H., Bos, J., and Pulman, S., editors, *Computing Meaning*, volume 4. Springer, Dordrecht.

[Bunt, 2015] Bunt, H. (2015). On the principles of semantic annotation. In *Proceedings of the 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-11)*, UK.

[Bunt, 2017] Bunt, H. (2017). Towards interoperable annotation of quantification. In *Proceedings of the 13th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-13)*, pages 84–95, Montpellier, France.

[Bunt et al., 1995] Bunt, H., Ahn, R., Beun, R.-J., Borghuis, T., and van Overveld, K. (1995). Multimodal cooperation with the denk system. In *International Conference on Cooperative Multimodal Communication*, pages 39–67. Springer.

[Bunt et al., 2010] Bunt, H., Alexandersson, J., Carletta, J., Choe, J.-W., Fang, A. C., Hasida, K., Lee, K., Petukhova, V., Popescu-Belis, A., Romary, L., et al. (2010). Towards an ISO standard for dialogue act annotation. In *Seventh conference on International Language Resources and Evaluation (LREC'10)*.

[Bunt et al., 2012a] Bunt, H., Alexandersson, J., Choe, J.-W., Fang, A. C., Hasida, K., Petukhova, V., Popescu-Belis, A., and Traum, D. R. (2012a). ISO 24617-2: A semantically-based standard for dialogue annotation. In *LREC*, pages 430–437. Citeseer.

[Bunt et al., 2013] Bunt, H., Fang, A., Liu, X., Cao, J., and Petukhova, V. (2013). Issues in the addition of ISO standard annotations to the Switchboard corpus. In *Workshop on Interoperable Semantic Annotation*, page 67.

[Bunt et al., 2007] Bunt, H., Keizer, S., and Morante, R. (2007). A computational model of grounding in dialogue. In *Proceedings of the Workshop in Discourse and Dialogue. Lecture Notes in Computer Science 4629*, pages 591–598, Antwerp, Belgium.

[Bunt et al., 2012b] Bunt, H., Kipp, M., and Petukhova, V. (2012b). Using DiAML and ANVIL for multimodal dialogue annotation. In *Proceedings 8th International Conference on Language Resources and Evaluation*, Istanbul, Turkey. ELRA, Paris.

[Bunt et al., 2017a] Bunt, H., Petukhova, V., and Fang, A. (2017a). Revisiting the ISO standard for dialogue act annotation. In *Proceedings of the 13th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-13)*, pages 37–51, Montpellier, France.

[Bunt et al., 2016] Bunt, H., Petukhova, V., Malchanau, A., A., F., and Wijnhoven, K. (2016). The DialogBank. In *Proceedings 9th International Conference on Language Resources and Evaluation (LREC 2016)*, Portoroz, Slovenia. ELRA, Paris.

[Bunt et al., 2017b] Bunt, H., Petukhova, V., Traum, D., and Alexandersson, J. (2017b). Dialogue act annotation with the iso 24617-2 standard. In *Multimodal interaction with W3C standards*, pages 109–135. Springer.

[Bunt and Prasad, 2016] Bunt, H. and Prasad, R. (2016). ISO DR-Core (ISO 24617-8): Core concepts for the annotation of discourse relations. In *Proceedings 12th Joint ACL-ISO Workshop on Interoperable Semantic Annotation*, pages 45–54, Portoroz, Slovenia.

[Cadilhac et al., 2013] Cadilhac, A., Asher, N., Benamara, F., and Lascarides, A. (2013). Grounding strategic conversation: Using negotiation dialogues to predict trades in a win-lose game. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 357–368.

[Caillou et al., 2017] Caillou, P., Gaudou, B., Grignard, A., Truong, C. Q., and Taillandier, P. (2017). A simple-to-use BDI architecture for agent-based modeling and simulation. In *Advances in Social Simulation 2015*, pages 15–28. Springer.

[Callejas et al., 2011] Callejas, Z., López-Cózar, R., Ábalos, N., and Griol, D. (2011). Affective conversational agents: The role of personality and emotion in spoken interactions. *Conversational Agents and Natural Language Interaction: Techniques and Effective Practices*, page 203.

[Carberry, 1990] Carberry, S. (1990). *Plan recognition in natural language dialogue. ACL-MIT Press Series in Natural Language Processing*. Bradford Books, MIT Press, Cambridge, Massachusetts.

[Card et al., 1980a] Card, S. K., Moran, T. P., and Newell, A. (1980a). Computer text-editing: An information-processing analysis of a routine cognitive skill. *Cognitive psychology*, 12(1):32–74.

[Card et al., 1980b] Card, S. K., Moran, T. P., and Newell, A. (1980b). The keystroke-level model for user performance time with interactive systems. *Communications of the ACM*, 23(7):396–410.

[Card et al., 1983] Card, S. K., Moran, T. P., and Newell, A. (1983). *The psychology of human-computer interaction*. Lawrence Erlbaum Associates.

[Carletta, 2006] Carletta, J. (2006). Announcing the AMI meeting corpus. *The ELRA Newsletter*, 11(1):3–5.

[Carlson, 1997] Carlson, R. (1997). *Experienced cognition*. Psychology Press.

[Chi et al., 2001] Chi, M., Siler, S., Jeong, H., Yamauchi, T., and Hausmann, R. (2001). Learning from human tutoring. *Cognitive Science*, 25(4):471–533.

[Chin et al., 1988] Chin, J., Diehl, V., and Norman, K. (1988). Development of an instrument measuring user satisfaction of the human-computer interface. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 213–218. ACM.

[Chipman et al., 2000] Chipman, S. F., Schraagen, J. M., and Shalin, V. L. (2000). Introduction to cognitive task analysis. *Cognitive task analysis*, pages 3–23.

[Chu-Carroll and Carberry, 2000] Chu-Carroll, J. and Carberry, S. (2000). Conflict resolution in collaborative planning dialogs. *International Journal of Human-Computer Studies*, 53(6):969–1015.

[Clark, 1998] Clark, A. (1998). *Being there: Putting brain, body, and world together again*. MIT press.

[Clark, 1996] Clark, H. (1996). *Using language*. Cambridge University Press.

[Clark and Krych, 2004] Clark, H. and Krych, M. (2004). Speaking while monitoring addressees for understanding. *Journal of memory and language*, 50(1):62–81.

[Clark and Schaefer, 1989] Clark, H. and Schaefer, E. (1989). Contributing to discourse. *Cognitive Science*, 13:259–294.

[Clark and Wilkes-Gibbs, 1986] Clark, H. and Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, 22(1):1–39.

[Cohen and Perrault, 1979] Cohen, P. and Perrault, C. (1979). Elements of a plan-based theory of speech acts. *Cognitive science*, 3(3):177–212.

[Cohn and Ekman, 2005] Cohn, J. and Ekman, P. (2005). Measuring facial action. *The new handbook of methods in nonverbal behavior research*, pages 9–64.

[Cooper, 2004] Cooper, R. (2004). A type-theoretic approach to information state update in issue-based dialogue management. In *Invited talk at CATALOG-04, 8th Workshop on the Semantics and Pragmatics of Dialogue*, Barcelona, Spain.

[Corbett and Chang, 1983] Corbett, A. and Chang, F. (1983). Pronoun disambiguation: Accessing potential antecedents. *Memory & Cognition*, 11(3):283–294.

[Core et al., 2014] Core, M., Lane, C., and Traum, D. (2014). Intelligent tutoring support for learners interacting with virtual humans. In Sottilare, R., Graesser, A., Hu, X., and Goldberg, B., editors, *Design Recommendations for Intelligent Tutoring Systems*, volume 2, pages 249–257. U.S. Army Research Laboratory, Orlando, FL, USA.

[Cox, 1958] Cox, A. (1958). The duty to bargain in good faith. *Harvard Law Review*, pages 1401–1442.

[Crangle and Suppes, 1994] Crangle, C. and Suppes, P. (1994). Language and learning for robots. In *CSLI Lecture notes*, volume 41. Centre for the Study of Language and Communication, Stanford, CA.

[Crismore et al., 1993] Crismore, A., Markkanen, R., and Steffensen, M. (1993). Metadiscourse in persuasive writing: A study of texts written by American and Finnish university students. *Written communication*, 10(1):39–71.

[Crook et al., 2010] Crook, P., Henderson, J., Lemon, O., and Liu, X. (2010). Combining POMDP approaches with ISU dialogue management. *Computer Speech and Language*.

[Csikszentmihalyi, 1997] Csikszentmihalyi, M. (1997). Flow and the psychology of discovery and invention. *HarperPerennial, New York*, 39.

[Dahan et al., 2002] Dahan, D., Tanenhaus, M., and Chambers, C. (2002). Accent and reference resolution in spoken-language comprehension. *Journal of Memory and Language*, 47(2):292–314.

[Dahlbäck et al., 1993] Dahlbäck, N., Jönsson, A., and Ahrenberg, L. (1993). Wizard of Oz studies - why and how. *Knowledge-based systems*, 6(4):258–266.

[Dahlbaeck and Jonsson, 1998] Dahlbaeck, N. and Jonsson, A. (1998). A coding manual for the Linköping dialogue model. Unpublished manuscript.

[Damasio, 2004] Damasio, A. (2004). Emotions and feelings: A neurobiological perspective.

[Danieli and Gerbino, 1995] Danieli, M. and Gerbino, E. (1995). Metrics for evaluating dialogue strategies in a spoken language system. In *Proceedings of the 1995 AAAI spring symposium on Empirical Methods in Discourse Interpretation and Generation*, volume 16, pages 34–39.

[Daubigney et al., 2012] Daubigney, L., Geist, M., Chandramohan, S., and Pietquin, O. (2012). A comprehensive reinforcement learning framework for dialogue management optimization. *IEEE Journal of Selected Topics in Signal Processing*, 6(8):891–902.

[Davidson, 1971] Davidson, D. (1971). I. agency. In Ausonio Marras, R. N. B. and Binkley, R. W., editors, *Agent, Action, and Reason*, pages 1–37. University of Toronto Press.

[De Dreu et al., 2000] De Dreu, C., Weingart, L., and Kwon, S. (2000). Influence of social motives on integrative negotiation: a meta-analytic review and test of two theories.

[Deci and Ryan, 2011] Deci, E. L. and Ryan, R. M. (2011). Self-determination theory. *Handbook of theories of social psychology*, 1(2011):416–433.

[Dede, 2009] Dede, C. (2009). Immersive interfaces for engagement and learning. *Science*, 323(5910):66–69.

[Dix, 2009] Dix, A. (2009). Human-computer interaction. In *Encyclopedia of database systems*, pages 1327–1331. Springer.

[Dodge et al., 2015] Dodge, J., Gane, A., Zhang, X., Bordes, A., Chopra, S., Miller, A., Szlam, A., and Weston, J. (2015). Evaluating prerequisite qualities for learning end-to-end dialog systems. *arXiv preprint arXiv:1511.06931*.

[Dornaika and Davoine, 2006] Dornaika, F. and Davoine, F. (2006). On appearance based face and facial action tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, 16(9):1107–1124.

[Dornaika and Raducanu, 2009] Dornaika, F. and Raducanu, B. (2009). Simultaneous 3D face pose and person-specific shape estimation from a single image using a holistic approach. In *Proceedins of the IEEE International Workshop on Applications of Computer Vision*.

[D'Souza, 2013] D'Souza, C. (2013). Debating: a catalyst to enhance learning skills and competencies. *Education+ Training*, 55(6):538–549.

[Dzikovska et al., 2011] Dzikovska, M. O., Moore, J. D., Steinhauser, N., and Campbell, G. (2011). Exploring user satisfaction in a tutorial dialogue system. In *Proceedings of the SIGDIAL 2011 Conference*, pages 162–172. Association for Computational Linguistics.

[Ebhotemhen et al., 2017] Ebhotemhen, E., Petukhova, V., and Klakow, D. (2017). Incremental dialogue act recognition: token- vs chunk-based classification. In *Proceedings of the 18th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Stockholm, Sweden.

[Efstathiou and Lemon, 2015] Efstathiou, I. and Lemon, O. (2015). Learning non-cooperative dialogue policies to beat opponent models: "The good, the bad and the ugly". *SEMDIAL 2015 goDIAL*, page 33.

[Ekman and Friesen, 1977] Ekman, P. and Friesen, W. (1977). *Facial action coding system*. Consulting Psychologists Press, Stanford University, Palo Alto.

[Ericsson and Simon, 1993] Ericsson, K. A. and Simon, H. A. (1993). *Protocol analysis: Verbal reports as data, Rev*. the MIT Press.

[Fillmore, 1977] Fillmore, C. J. (1977). The case for case reopened. *Syntax and semantics*, 8(1977):59–82.

[Fischer, 2001] Fischer, G. (2001). User modeling in human–computer interaction. *User modeling and user-adapted interaction*, 11(1-2):65–86.

[Fisher and Ury, 1981] Fisher, R. and Ury, W. (1981). *Getting to yes: Negotiating agreement without giving in*. Harmondsworth, Middlesex: Penguin.

[Frampton and Lemon, 2009] Frampton, M. and Lemon, O. (2009). Recent research advances in reinforcement learning in spoken dialogue systems. *The Knowledge Engineering Review*, 24(4):375–408.

[Fraser, 1998] Fraser, N. (1998). Assessment of interactive systems. In *Handbook of standards and resources for spoken language systems*, pages 564–615. Mouton de Gruyter.

[Freeman, 2011] Freeman, J. (2011). Argument structure: representation and theory. In *Argumentation Library*, volume 18. Springer, Berlin.

[Friedman, 2002] Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4):367–378.

[Gales and Young, 2007] Gales, M. and Young, S. (2007). The application of Hidden Markov Models in speech recognition. *Foundations and Trends in Signal Processing*, 1:195–304.

[Gama, 2004] Gama, C. (2004). Metacognition in interactive learning environments: The reflection assistant model. In Lester, J. C., Vicario, R. M., and Paraguacu, F., editors, *Intelligent Tutoring Systems: 7th International Conference, ITS 2004, Maceió, Alagoas, Brazil, August 30 - September 3, 2004*, volume 3220 of *Lecture Notes in Computer Science*, pages 668–677. Springer.

[Geertzen et al., 2004] Geertzen, J., Girard, Y., Morante, R., Van der Sluis, I., Van Dam, H., Suijkerbuijk, B., Van der Werf, R., and Bunt, H. (2004). The DIAMOND project. In *Proceedings of the 8th Workshop on the Semantics and Pragmatics of Dialogue, Catalog'04*.

[Geertzen et al., 2007] Geertzen, J., Petukhova, V., and Bunt, H. (2007). A multidimensional approach to utterance segmentation and dialogue act classification. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, pages 140–149, Antwerp, Belgium. Association for Computational Linguistics.

[Georgila et al., 2018a] Georgila, K., Gordon, C., Choi, H., Boberg, J., Jeon, H., and Traum, D. (2018a). Toward low-cost automated evaluation metrics for Internet of Things dialogues. In *Proceedings Ninth International Workshop on Spoken Dialogue Systems Technology (ISWDS 2018)*, Singapore.

[Georgila et al., 2018b] Georgila, K., Gordon, C., Choi, H., Boberg, J., Jeon, H., and Traum, D. (2018b). Toward low-cost automated evaluation metrics for internet of things dialogues. In *Proceedings Ninth International Workshop on Spoken Dialogue Systems Technology (ISWDS 2018)*, Singapore.

[Georgila and Traum, 2011] Georgila, K. and Traum, D. (2011). Reinforcement learning of argumentation dialogue policies in negotiation. In *Twelfth Annual Conference of the International Speech Communication Association*.

[Georgila et al., 2010] Georgila, K., Wolters, M., and Moore, J. (2010). Learning dialogue strategies from older and younger simulated users. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 103–106. Association for Computational Linguistics.

[Ginzburg, 1996] Ginzburg, J. (1996). Dynamics and the semantics of dialogue. *Logic, language and computation*, 1.

[Ginzburg, 1998] Ginzburg, J. (1998). Clarifying utterances. In *Proceedings of the Twente Workshop on the Formal Semantics and Pragmatics of Dialogues*, pages 11–30, Enschede, The Netherlands.

[Gnjatovic and Rösner, 2008] Gnjatovic, M. and Rösner, D. (2008). *Emotion adaptive dialogue management in human-machine interaction*. Citeseer.

[Gonzalez and Lebiere, 2005] Gonzalez, C. and Lebiere, C. (2005). Instance-based cognitive models of decision-making. In Zizzo, D. and Courakis, A., editors, *Transfer of knowledge in economic decision making*. Macmillan.

[Gratch et al., 2007] Gratch, J., Wang, N., Gerten, J., Fast, E., and Duffy, R. (2007). Creating rapport with virtual agents. In *International Workshop on Intelligent Virtual Agents*, pages 125–138. Springer.

[Greenbaum and Kyng, 1992] Greenbaum, J. and Kyng, M. (1992). *Design at work: Cooperative design of computer systems*. L. Erlbaum Associates Inc.

[Grice, 1975] Grice, H. (1975). Logic and conversation. In Cole, P. and Morgan, J., editors, *Syntax and Semantics: Speech Acts*, volume 3, pages 43–58. Academic Press, New York.

[Grice and Savino, 2003] Grice, M. and Savino, M. (2003). Map Tasks in Italian: asking questions about given, accessible and new information. *Catalan journal of linguistics*, 2:153–180.

[Gross, 2016] Gross, R. (2016). *Psychology: The science of mind & behaviour*. Hodder Education.

[Grosz and Sidner, 1986] Grosz, B. J. and Sidner, C. L. (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12:175–204.

[Grosz and Sidner, 1990] Grosz, B. J. and Sidner, C. L. (1990). Plans for discourse. In Cohen, P., Morgan, J., and Pollack, M., editors, *Intentions in Communication*, pages 417–444. MIT Press, Cambridge, Massachusetts.

[Guhe and Lascarides, 2014] Guhe, M. and Lascarides, A. (2014). Persuasion in complex games. *Proceedings of the 18th Workshop on the Semantics and Pragmatics of Dialogue (SEMDIAL 2014 - DialWatt)*, page 62.

[Gunzelmann et al., 2009] Gunzelmann, G., Moore, L., Gluck, K., Van Dongen, H., and Dinges, D. (2009). Examining sources of individual variation in sustained attention. In *Proceedings of the thirty-first annual meeting of the cognitive science society*, pages 608–613.

[Haider et al., 2017] Haider, F., Luz, S., and Campbell, N. (2017). Data collection and synchronisation: Towards a multiperspective multimodal dialogue system with metacognitive abilities. In *Dialogues with Social Robots*, pages 245–256. Springer.

[Harley et al., 2013] Harley, J., Bouchet, F., and Azevedo, R. (2013). Aligning and comparing data on emotions experienced during learning with metaTutor. In *International Conference on Artificial Intelligence in Education*, pages 61–70. Springer.

[Hastie et al., 2009] Hastie, T., Rosset, S., Zhu, J., and Zou, H. (2009). Multi-class adaboost. *Statistics and its Interface*, 2(3):349–360.

[Henderson et al., 2008] Henderson, J., Lemon, O., and Georgila, K. (2008). Hybrid reinforcement/supervised learning of dialogue policies from fixed data sets. *Computational Linguistics*, 34(4):487–511.

[Henderson et al., 2014a] Henderson, M., Thomson, B., and Williams, J. (2014a). The second dialog state tracking challenge. In *SIGDIAL Conference*, pages 263–272.

[Henderson et al., 2013] Henderson, M., Thomson, B., and Young, S. (2013). Deep neural network approach for the dialog state tracking challenge. In *SIGDIAL Conference*, pages 467–471.

[Henderson et al., 2014b] Henderson, M., Thomson, B., and Young, S. (2014b). Robust dialog state tracking using delexicalised recurrent neural networks and unsupervised adaptation. In *Spoken Language Technology Workshop (SLT), 2014 IEEE*, pages 360–365. IEEE.

[Henderson et al., 2014c] Henderson, M., Thomson, B., and Young, S. (2014c). Word-based dialog state tracking with recurrent neural networks. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 292–299.

[Hindriks et al., 2007] Hindriks, K., Jonker, C., and Tykhonov, D. (2007). Analysis of negotiation dynamics. In *International Workshop on Cooperative Information Agents*, pages 27–35. Springer.

[Hirschberg, 2002] Hirschberg, J. (2002). The pragmatics of intonational meaning. In *Speech Prosody 2002, International Conference*, pages 65–68, Aix-en-Provence, France. Laboratoire Parole et Langage.

[Hodgkinson and Crawshaw, 1985] Hodgkinson, G. and Crawshaw, C. (1985). Hierarchical task analysis for ergonomics research: An application of the method to the design and evaluation of sound mixing consoles. *Applied Ergonomics*, 16(4):289–299.

[Hone and Graham, 2001] Hone, K. S. and Graham, R. (2001). Subjective assessment of speech-system interface usability. In *Seventh European Conference on Speech Communication and Technology*.

[Hovy and Maier, 1995] Hovy, E. and Maier, E. (1995). Parsimonious of profligate: how many and which discourse structure relations? unpublished manuscript.

[Hughes et al., 2013] Hughes, T., Flatt, J., Fu, B., Chang, C.-C., and Ganguli, M. (2013). Engagement in social activities and progression from mild to severe cognitive impairment: the myhat study. *International psychogeriatrics*, 25(04):587–595.

[Ickes et al., 2006] Ickes, W., Holloway, R., Stinson, L., and Hoodenpyle, T. (2006). Self-monitoring in social interaction: The centrality of self-affect. *Journal of personality*, 74(3):659–684.

[Ide and Pustejovsky, 2017] Ide, N. and Pustejovsky, J. (2017). *Handbook of Linguistic Annotation*. Springer.

[Ide and Romary, 2004] Ide, N. and Romary, L. (2004). International standard for a linguistic annotation framework. *Natural language engineering*, 10(3-4):211–225.

[Isard, 1975] Isard, S. (1975). Changing the context. *Formal Semantics of Natural Language. Cambridge University Press, Cambridge, England*, pages 287–296.

[ISO, 2006] ISO (2006). *TEI-ISO 24610-1:2006 Language resource management: Feature structures, Part 1: Feature structure representation*. ISO, Geneve.

[ISO, 2012] ISO (2012). *Language resource management – Semantic annotation framework – Part 2: Dialogue acts. ISO 24617-2*. ISO Central Secretariat, Geneva.

[ISO, 2016] ISO (2016). *Language resource management – Semantic annotation framework – Part 6: Principles of Semantic Annotation. ISO 24617-6*. ISO Central Secretariat, Geneva.

[Jäger, 2004] Jäger, R. (2004). Konstruktion einer ratingskala mit smilies als symbolische marken. *Diagnostica*, 50(1):31–38.

[Janarthanam and Lemon, 2009] Janarthanam, S. and Lemon, O. (2009). A two-tier user simulation model for reinforcement learning of adaptive referring expression generation policies. In *Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 120–123. Association for Computational Linguistics.

[Johnson, 2009] Johnson, S. (2009). *Winning Debates: A Guide to Debating in the Style of the World Universities Debating Championships*. G - Reference, Information and Interdisciplinary Subjects Series. International Debate Education Association, Brussels, Belgium.

[Jonassen et al., 1989] Jonassen, D. H., Hannum, W. H., and Tessmer, M. (1989). *Handbook of task analysis procedures.* Praeger Publishers.

[Jurafsky et al., 1997] Jurafsky, D., Schriberg, E., and Biasca, D. (1997). Switchboard SWBD-DAMSL shallow-discourse-function annotation: Coders manual.

[Kaptelinin and Czerwinski, 2007] Kaptelinin, V. and Czerwinski, M. (2007). *Beyond the desktop metaphor: designing integrated digital work environments*, volume 1. The MIT Press.

[Keizer et al., 2011] Keizer, S., Bunt, H., and Petukhova, V. (2011). Multidimensional dialogue management. In van den Bosch, A. and Bouma, G., editors, *IMIX book*. Springer.

[Keller, 2009] Keller, J. M. (2009). *Motivational design for learning and performance: The ARCS model approach*. Springer Science & Business Media.

[Kelley and Stahelski, 1970] Kelley, H. and Stahelski, A. (1970). Social interaction basis of cooperators' and competitors' beliefs about others. *Journal of personality and social psychology*, 16(1):66.

[Kiefer et al., 2008] Kiefer, M., Sim, E.-J., Herrnberger, B., Grothe, J., and Hoenig, K. (2008). The sound of concepts: four markers for a link between auditory and conceptual brain systems. *Journal of Neuroscience*, 28(47):12224–12230.

[Kieras, 1988] Kieras, D. E. (1988). Towards a practical GOMS model methodology for user interface design. In *Handbook of human-computer interaction*, pages 135–157. Elsevier.

[Kim et al., 2015] Kim, S., D'Haro, L., Banchs, R., Williams, J., and Henderson, M. (2015). Dialog state tracking challenge 4.

[Kooijmans et al., 2007] Kooijmans, T., Kanda, T., Bartneck, C., Ishiguro, H., and Hagita, N. (2007). Accelerating robot development through integral analysis of human–robot interaction. *IEEE Transactions on Robotics*, 23(5):1001–1012.

[Koryzis et al., 2016] Koryzis, D., Svolopoulos, V., and Spiliotopoulos, D. (2016). Metalogue: A multimodal learning journey. In *Proceedings of the 9th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, page 48, Corfu, Island, Greece. ACM.

[Krahmer and Swerts, 2007] Krahmer, E. and Swerts, M. (2007). The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception. *Journal of Memory and Language*, 57(3):396–414.

[Kreibig et al., 2007] Kreibig, S., Wilhelm, F., Roth, W., and Gross, J. (2007). Cardiovascular, electrodermal, and respiratory response patterns to fear-and sadness-inducing films. *Psychophysiology*, 44(5):787–806.

[Kring and Sloan, 2007] Kring, A. and Sloan, D. (2007). The facial expression coding system (faces): development, validation, and utility. *Psychological assessment*, 19(2):210.

[Lafferty et al., 2001] Lafferty, J., McCallum, A., Pereira, F., et al. (2001). Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML '01*, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[Lapina and Petukhova, 2017] Lapina, V. and Petukhova, V. (2017). Classification of modal meaning in negotiation dialogues. In *Proceedings of the 13th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-13)*, pages 59–70, Montpellier, France.

[Larsen and Prizmic-Larsen, 2006] Larsen, R. and Prizmic-Larsen, Z. (2006). Measuring emotions: Implications of a multimethod perspective. In Eid, M. and Diener, E., editors, *Handbook of multimethod measurement in psychology*, pages 337–351. American Psychological Association.

[Larsson, 2002] Larsson, S. (2002). *Issue-based dialogue management. PhD thesis*. Göteborg University, Göteborg, Sweden.

[Larsson et al., 2000] Larsson, S., Ljunglöf, P., Cooper, R., Engdahl, E., and Ericsson, S. (2000). GoDis: an accommodating dialogue system. In *Proceedings ANLP/NAACL 2000 Workshop on Conversational Systems, Seattle, Washington*, pages 7–10.

[Larsson and Traum, 2000] Larsson, S. and Traum, D. (2000). Information state and dialogue management in the Trindi dialogue move engine toolkit. *Natural Language Engineering*, 6(3-4):323–340.

[Lauria et al., 2001] Lauria, S., Bugmann, G., Kyriacou, T., Bos, J., and Klein, A. (2001). Training personal robots using natural language instruction. *IEEE Intelligent systems*, 16(5):38–45.

[Lax and Sebenius, 1992] Lax, D. and Sebenius, J. (1992). The manager as negotiator: The negotiators dilemma: Creating and claiming value. *Dispute resolution*, 2:49–62.

[Lebiere and Anderson, 1998] Lebiere, C. and Anderson, J. R. (1998). Cognitive arithmetic. *The atomic components of thought*, pages 297–342.

[Lebiere et al., 1998] Lebiere, C., Wallach, D., and Taatgen, N. (1998). Implicit and explicit learning in ACT-R. In *Proceedings of the 2nd European Conference on Cognitive Modelling (ECCM 98)*. University of Nottingham Press.

[Lebiere et al., 2000] Lebiere, C., Wallach, D., and West, R. (2000). A memory-based account of the prisoner's dilemma and other 2x2 games. In *Proceedings of International Conference on Cognitive Modeling*, pages 185–193.

[Lee et al., 2015] Lee, H., Betts, S., and Anderson, J. (2015). Learning problem-solving rules as search through a hypothesis space. *Cognitive science*.

[Lee et al., 2017] Lee, K., Zhao, T., Du, Y., Cai, E., Lu, A., Pincus, E., Traum, D., Ultes, S., Rojas Barahona, L., Gasic, M., Young, S., and Eskenazi, M. (2017). DialPort, gone live: An update after a year of development. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 170–173, Saarbrücken, Germany. Association for Computational Linguistics.

[Lemon et al., 2001] Lemon, O., Bracy, A., Gruenstein, A., and Peters, S. (2001). The WITAS multi-modal dialogue system i. In *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech 2001)*, pages 1559–1562.

[Lemon et al., 2003] Lemon, O., Cavedon, L., and Kelly, B. (2003). Managing dialogue interaction: A multi-layered approach. In *Proceedings of the 4th SIGdial workshop on Discourse and Dialogue*.

[Lemon et al., 2006] Lemon, O., Georgila, K., and Henderson, J. (2006). Evaluating effectiveness and portability of reinforcement learned dialogue strategies with real users: the talk towninfo evaluation. In *Spoken Language Technology Workshop, 2006. IEEE*, pages 178–181. IEEE.

[Lessiter et al., 2001] Lessiter, J., Freeman, J., Keogh, E., and Davidoff, J. (2001). A cross-media presence questionnaire: The ITC-sense of presence inventory. *Presence*, 10(3):282–297.

[Lewin, 1998] Lewin, I. (1998). The autoroute dialogue demonstrator. Technical Report CRC-073, SRI Cambridge Computer Science Research Centre.

[Lewis, 1991] Lewis, J. R. (1991). Psychometric evaluation of an after-scenario questionnaire for computer usability studies: the asq. *ACM Sigchi Bulletin*, 23(1):78–81.

[Lewis et al., 2017] Lewis, M., Yarats, D., Dauphin, Y. N., Parikh, D., and Batra, D. (2017). Deal or no deal? end-to-end learning for negotiation dialogues. *arXiv preprint arXiv:1706.05125*.

[Lim et al., 2012] Lim, M., Dias, J., Aylett, R., and Paiva, A. (2012). Creating adaptive affective autonomous NPCs. *Autonomous Agents and Multi-Agent Systems*, 24(2):287–311.

[Linek et al., 2008] Linek, S., Marte, B., and Albert, D. (2008). The differential use and effective combination of questionnaires and logfiles. In *Computer-based Knowledge & Skill Assessment and Feedback in Learning settings (CAF), Proceedings of the International Conference on Interactive Computer Aided Learning (ICL), 24th to 26th September*.

[Litman and Silliman, 2004] Litman, D. and Silliman, S. (2004). ITSPOKE: An intelligent tutoring spoken dialogue system. In *Demonstration papers at HLT-NAACL 2004*, pages 5–8. Association for Computational Linguistics.

[Liu et al., 2016] Liu, C.-W., Lowe, R., Serban, I. V., Noseworthy, M., Charlin, L., and Pineau, J. (2016). How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023*.

[Logan, 1988] Logan, G. (1988). Toward an instance theory of automatization. *Psychological review*, 95(4):492.

[López-Cózar et al., 2006] López-Cózar, R., Callejas, Z., and McTear, M. (2006). Testing the performance of spoken dialogue systems by means of an artificially simulated user. *Artificial Intelligence Review*, 26(4):291–323.

[López-Cózar et al., 2009] López-Cózar, R., Espejo, G., Callejas, Z., Gutiérrez, A., and Griol, D. (2009). Assessment of spoken dialogue systems by simulating different levels of user cooperativeness. *Methods*, 1(8):9.

[Mager, 1997] Mager, R. F. (1997). Making instruction work or skillbloomers: A stepbystep guide to designing and developing instruction that works. *Atlanta, GA: The Center for Effective Performance*.

[Mairesse and Walker, 2005] Mairesse, F. and Walker, M. (2005). Learning to personalize spoken generation for dialogue systems. In *INTERSPEECH*, pages 1881–1884.

[Malchanau et al., 2018a] Malchanau, A., Petukhova, V., and Bunt, H. (2018a). Multimodal dialogue system evaluation: a case study applying usability standards. In *Proceedings Ninth International Workshop on Spoken Dialogue Systems Technology (ISWDS 2018)*, Singapore.

[Malchanau et al., 2018b] Malchanau, A., Petukhova, V., and Bunt, H. (2018b). Towards continuous dialogue corpus creation: Writing to corpus and generating from it. In *Proceedings of the 11th International Conference on Language Resources and Evaluation(LREC 2018)*, Myiazaki, Japan.

[Malchanau et al., 2019] Malchanau, A., Petukhova, V., and Bunt, H. (2019). Towards integration of cognitive models in dialogue management: Designing the virtual negotiation coach application. *Dialogue & Discourse*, 9(2):35–79.

[Malchanau et al., 2015] Malchanau, A., V., P., H., B., and D., K. (2015). Multidimensional dialogue management for tutoring systems. In *Proceedings of the 7th Language and Technology Conference (LTC 2015)*, Poznan, Poland.

[Marewski and Link, 2014] Marewski, J. and Link, D. (2014). Strategy selection: An introduction to the modeling challenge. *Wiley Interdisciplinary Reviews: Cognitive Science*, 5(1):39–59.

[Martin et al., 1999] Martin, D., Cheyer, A., and Moran, D. (1999). The Open Agent Architecture: A framework for building distributed software systems. *Applied Artificial Intelligence*, 13(1-2):91–128.

[Mauss and Robinson, 2009] Mauss, I. and Robinson, M. (2009). Measures of emotion: A review. *Cognition and emotion*, 23(2):209–237.

[Means, 1993] Means, B. (1993). Cognitive task analysis as a basis for instructional design. *Cognitive science foundations of instruction*, pages 97–118.

[Meijering et al., 2014] Meijering, B., Taatgen, N. A., van Rijn, H., and Verbrugge, R. (2014). Modeling inference of mental states: As simple as possible, as complex as necessary. *Interaction Studies*, 15(3):455–477.

[Meijering et al., 2012] Meijering, B., Van Rijn, H., Taatgen, N., and Verbrugge, R. (2012). What eye movements can tell about theory of mind in a strategic game. *PloS one*, 7(9):e45961.

[Möller, 2004] Möller, S. (2004). *Quality of telephone-based spoken dialogue systems*. Springer Science & Business Media.

[Moore et al., 2005] Moore, D., Cheng, Y., McGrath, P., and Powell, N. (2005). Collaborative virtual environment technology for people with autism. *Journal of the Hammill Institute on Disabilities*, 20(4):231243.

[Mory, 2004] Mory, E. (2004). Feedback research revisited. *Handbook of research on educational communications and technology*, 2:745–783.

[Nasoz and Lisetti, 2007] Nasoz, F. and Lisetti, C. (2007). Affective user modeling for adaptive intelligent user interfaces. In *Human-Computer Interaction. HCI Intelligent Multimodal Interaction Environments*, pages 421–430. Springer.

[Nass et al., 2005] Nass, C., Jonsson, I.-M., Harris, H., Reaves, B., Endo, J., Brave, S., and Takayama, L. (2005). Improving automotive safety by pairing driver emotion and car voice emotion. In *CHI'05 Extended Abstracts on Human Factors in Computing Systems*, pages 1973–1976. ACM.

[Nass and Lee, 2000] Nass, C. and Lee, K. (2000). Does computer-generated speech manifest personality? An experimental test of similarity-attraction. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 329–336. ACM.

[Niedenthal et al., 2005] Niedenthal, P. M., Barsalou, L. W., Winkielman, P., Krauth-Gruber, S., and Ric, F. (2005). Embodiment in attitudes, social perception, and emotion. *Personality and social psychology review*, 9(3):184–211.

[Nielsen, 2012] Nielsen, J. (2012). User satisfaction vs. performance metrics. *Nielsen Norman Group*.

[Niewiadomski et al., 2008] Niewiadomski, R., Ochs, M., and Pelachaud, C. (2008). Expressions of empathy in ECAs. In *International Workshop on Intelligent Virtual Agents*, pages 37–44. Springer.

[Nijboer et al., 2016] Nijboer, M., Borst, J., van Rijn, H., and Taatgen, N. (2016). Contrasting single and multi-component working-memory systems in dual tasking. *Cognitive psychology*, 86:1–26.

[Nilsson and Fikes, 1970] Nilsson, N. J. and Fikes, R. E. (1970). STRIPS: A new approach to the application of theorem proving to problem solving. Technical report, SRI Internaltional, Menlo Park, CA, Artificial Intelligence Center.

[Nir, 1988] Nir, R. (1988). Electoral rhetoric in Israel - the television debates. a study in political discourse. *Language Learning*, 38:2:187–208.

[Nisimura et al., 2006] Nisimura, R., Omae, S., Kawahara, H., and Irino, T. (2006). Analyzing dialogue data for real-world emotional speech classification. In *Ninth International Conference on Spoken Language Processing*.

[Norman, 1993] Norman, D. A. (1993). Things that make us smart: Defending human attributes in the age of the machine.

[Nouri et al., 2017] Nouri, E., Georgila, K., and Traum, D. (2017). Culture-specific models of negotiation for virtual characters: multi-attribute decision-making based on culture-specific values. *AI & society*, 32(1):51–63.

[Novák-Tót et al., 2017] Novák-Tót, E., Niebuhr, O., and Chen, A. (2017). A gender bias in the acoustic-melodic features of charismatic speech? In *Proceedings of the 18th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2248–2252, Stockholm, Sweden. International Speech Communication Association (ISCA), Baixas, France.

[Ormerod and Shepherd, 2003] Ormerod, T. C. and Shepherd, A. (2003). Using task analysis for information requirements specification: the sub-goal template (sgt) method. *The handbook of task analysis for human-computer interaction*, page 347.

[Paiva et al., 2004] Paiva, A., Dias, J., Sobral, D., Aylett, R., Sobreperez, P., Woods, S., Zoll, C., and Hall, L. (2004). Caring for agents and agents that care: Building empathic relations with synthetic agents. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems-Volume 1*, pages 194–201. IEEE Computer Society.

[Papineni et al., 2002] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

[Pejčić, 2014] Pejčić, A. (2014). Intonational characteristics of persuasiveness in Serbian and English political debates. *Nouveaux Cahiers de Linguistique Française*, 31:141–151.

[Pekrun and Perry, 2014] Pekrun, R. and Perry, R. (2014). Control-value theory of achievement emotions. *International handbook of emotions in education*, pages 120–141.

[Pelachaud et al., 2002] Pelachaud, C., Carofiglio, V., De Carolis, B., de Rosis, F., and Poggi, I. (2002). Embodied contextual agent in information delivering application. In *Proceedings of the first international joint conference on Autonomous agents and multiagent systems: part 2*, pages 758–765. ACM.

[Peldszus and Stede, 2013] Peldszus, A. and Stede, M. (2013). From argument diagrams to argumentation mining in texts: a survey. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 7(1):1–31.

[Peltason and Wrede, 2011] Peltason, J. and Wrede, B. (2011). The curious robot as a case-study for comparing dialog systems. *AI magazine*, 32(4):85–99.

[Petukhova, 2011] Petukhova, V. (2011). *Multidimensional Dialogue Modelling. PhD dissertation*. Tilburg University, The Netherlands.

[Petukhova, 2014] Petukhova, V. (2014). Understanding questions and finding answers: semantic relation annotation to compute the expected answer type. In *Proceedings 10th Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation*, page 44.

[Petukhova and Bunt, 2010] Petukhova, V. and Bunt, H. (2010). Towards an integrated scheme for semantic annotation of multimodal dialogue data. In *Proceedings of LREC 2010*, pages 2556 – 2563, Malta.

[Petukhova et al., 2017a] Petukhova, V., Bunt, H., and Malchanau, A. (2017a). Computing negotiation update semantics in multi-issue bargaining dialogues. In *Proceedings of the SemDial 2017 (SaarDial) Workshop on the Semantics and Pragmatics of Dialogue*, Saarbrücken, Germany.

[Petukhova et al., 2015a] Petukhova, V., Bunt, H., Malchanau, A., and Aruchamy, R. (2015a). Experimenting with grounding strategies in dialogue. In *Proceedings of the GoDial 2015 Workshop on the Semantics and Pragmatics of Dialogue*, Goteborg, Sweden.

[Petukhova et al., 2015b] Petukhova, V., Bunt, H., Malchanau, A., and Aruchamy, R. (2015b). Experimenting with grounding strategies in dialogue. In *Proceedings of the GoDial 2015 Workshop on the Semantics and Pragmatics of Dialogue*, Goteborg, Sweden.

[Petukhova et al., 2014a] Petukhova, V., Gropp, M., Klakow, D., Schmidt, A., Eigner, G., Topf, M., Srb, S., Motlicek, P., Potard, B., Dines, J., et al. (2014a). The DBOX corpus collection of spoken human-human and human-machine dialogues. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 14)*. European Language Resources Association (ELRA).

[Petukhova et al., 2014b] Petukhova, V., Malchanau, A., and Bunt, H. (2014b). Interoperability of dialogue corpora through ISO 24617-2-based querying. In *Proceedings of the LREC 2014*, Iceland.

[Petukhova et al., 2015c] Petukhova, V., Malchanau, A., and Bunt, H. (2015c). Modelling argumentative behaviour in parliamentary debates. In *Proceedings of the 15th Workshop on Computational Models of Natural Argument, Principles and Practice of Multi-Agent Systmes Conference (PRIMA 2015)*, Bertinoro, Italy.

[Petukhova et al., 2016a] Petukhova, V., Malchanau, A., and Bunt, H. (2016a). Modelling argumentative behaviour in parliamentary debates: data collection, analysis and test case. In Baldoni, M., Baroglio, C., Bex, F., Grasso, F., Green, N., Namazi-Rad, M.-R.and Numao, M., and Suarez, M., editors, *Principles and Practice of Multi-Agent Systems. Lecture Notes in Artificial Intelligence*, pages 26–46. Springer, Berlin.

[Petukhova et al., 2018] Petukhova, V., Malchanau, A., Oualil, Y., Klakow, D., Luz, S., Haider, F., Campbell, N., Koryzis, D., Spiliotopoulos, D., Albert, P., Linz, N., and Alexanderssons, J. (2018). The Metalogue Debate Trainee Corpus: Data collection and annotations. In *Proceedings of the 11th International Conference on Language Resources and Evaluation(LREC 2018)*, Myiazaki, Japan.

[Petukhova et al., 2017b] Petukhova, V., Mayer, T., Malchanau, A., and Bunt, H. (2017b). Virtual debate coach design: Assessing multimodal argumentation performance. In *Proceedings of the 2017 ACM on International Conference on Multimodal Interaction (ICMI 2017)*, Glasgow, UK. ACM.

[Petukhova et al., 2017c] Petukhova, V., Raju, M., and Bunt, H. (2017c). Multimodal markers of persuasive speech : designing a virtual debate coach. In *Proceedings of*

the 18th Annual Conference of the International Speech Communication Association (INTERSPEECH), pages 142–146, Stockholm, Sweden. International Speech Communication Association (ISCA), Baixas, France.

[Petukhova et al., 2016b] Petukhova, V., Stevens, C., de Weerd, H., Taatgen, N., Cnossen, F., and Malchanau, A. (2016b). Modelling multi-issue bargaining dialogues: Data collection, annotation design and corpus. In *Proceedings of the 9th LREC 2016*. ELRA, Paris.

[Piso, 1981] Piso, E. (1981). Task analysis for process-control tasks: The method of Annett et al. applied. *Journal of Occupational and Organizational Psychology*, 54(4):247–254.

[Poesio and Traum, 1998] Poesio, M. and Traum, D. (1998). Towards an axiomatization of dialogue acts. In *Proceedings of the Twente Workshop on the Formal Semantics and Pragmatics of Dialogues*, pages 207–222.

[Pollack, 1992] Pollack, M. E. (1992). The uses of plans. *Artificial Intelligence*, 57(1):43–68.

[Povey, 2011] Povey, D. (2011). The Kaldi Speech Recognition Toolkit. In *Proceedings of the 2011 IEEE Workshop on Automatic Speech Recognition and Understanding*, Big Island, HI, US. IEEE Signal Processing Society.

[Prasad et al., 2008] Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., and Webber, B. (2008). The Penn Discourse Treebank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Maroc.

[Premack and Woodruff, 1978] Premack, D. and Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain sciences*, 1(04):515–526.

[Pustejovsky et al., 2017] Pustejovsky, J., Bunt, H., and Zaenen, A. (2017). Designing annotation schemes: From theory to model. In *Handbook of Linguistic Annotation*, pages 21–72. Springer.

[Raiffa et al., 2002] Raiffa, H., Richardson, J., and Metcalfe, D. (2002). *Negotiation analysis: The science and art of collaborative decision making*. Harvard University Press.

[Ramanand et al., 2003] Ramanand, P., Sreenivasan, R., and Nampoori, V. (2003). Complexity of brain dynamics inferred from the sample entropy analysis of electroencephalogram. *Proceedings of the National Conferece on Nonlinear Systems & Dynamics (NCNSD'03)*, pages 205–208.

[Reinecke, 2003] Reinecke, A. (2003). Designing commercial applications with life-like characters. *Lecture notes in computer science*, pages 181–181.

[Renals et al., 2014] Renals, S., Carletta, J., Edwards, K., Bourlard, H., Garner, P., Popescu-Belis, A., Klakow, D., Girenko, A., Petukova, V., Wacker, P., et al. (2014).

Rockit: roadmap for conversational interaction technologies. In *Proceedings of the 2014 Workshop on Roadmapping the Future of Multimodal Interaction Research including Business Opportunities and Challenges*, pages 39–42. ACM.

[Repp and Drenhaus, 2015] Repp, S. and Drenhaus, H. (2015). Intonation influences processing and recall of left-dislocation sentences by indicating topic vs. focus status of dislocated referent. *Language, Cognition and Neuroscience*, 30(3):324–346.

[Rich and Sidner, 1998] Rich, C. and Sidner, C. (1998). COLLAGEN: A collaboration manager for software interface agents. *User Modeling and User-Adapted Interaction*, 8:3.

[Riedl and Stern, 2006] Riedl, M. and Stern, A. (2006). Believable agents and intelligent story adaptation for interactive storytelling. In *International Conference on Technologies for Interactive Digital Storytelling and Entertainment*, pages 1–12. Springer.

[Ritter et al., 2007] Ritter, S., Anderson, J., Koedinger, K., and Corbett, A. (2007). Cognitive Tutor: Applied research in mathematics education. *Psychonomic bulletin & review*, 14(2):249–255.

[Roll et al., 2007] Roll, I., Aleven, V., McLaren, B., and Koedinger, K. (2007). Can help seeking be tutored? Searching for the secret sauce of metacognitive tutoring. In Luckin, R., Koedinger, K., and Greer, J., editors, *Artificial Intelligence in Education: Building Technology Rich Learning Contexts that Work*, volume 158 of *Frontiers in Artificial Intelligence and Applications*, pages 203–210. IOS Press.

[Root and Draper, 1983] Root, R. W. and Draper, S. (1983). Questionnaires as a software evaluation tool. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 83–87. ACM.

[Rosenberg and Hirschberg, 2009] Rosenberg, A. and Hirschberg, J. (2009). Charisma perception from text and speech. *Speech Communication*, 51.7:640–655.

[Rosenfeld, 2000] Rosenfeld, R. (2000). Two decades of statistical language modeling: where do we go from here? *Proceedings of the IEEE*, 88(8):1270–1278.

[Rouvier et al., 2013] Rouvier, M., Dupuy, G., Gay, P., Khoury, E., Merlin, T., and Meignier, S. (2013). An open-source state-of-the-art toolbox for broadcast news diarization. In *INTERSPEECH-2013*, pages 1477–1481.

[Rus et al., 2009] Rus, V., Lintean, M., and Azevedo, R. (2009). Automatic detection of student mental models during prior knowledge activation in MetaTutor. *International Working Group on Educational Data Mining*.

[Ruttkay and Pelachaud, 2004] Ruttkay, Z. and Pelachaud, C. (2004). *From brows to trust: evaluating embodied conversational agents*, volume 7. Springer Science & Business Media.

[Sadek, 1991] Sadek, M. (1991). Dialogue acts are rational plans. In *Proceedings of the ESCA/ETRW Workshop on the Structure of Multimodal Dialogue*, pages 19–48, Maratea, Italy.

[Sadowski and Stanney, 2002] Sadowski, W. and Stanney, K. (2002). Presence in virtual environments. In Hale, K. and Stanney, K., editors, *Handbook of Virtual Environments: Design, Implementation, and Applications*, pages 791—806. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US.

[Sali et al., 2010] Sali, S., Wardrip-Fruin, N., Dow, S., Mateas, M., Kurniawan, S., Reed, A., and Liu, R. (2010). Playing with words: from intuition to evaluation of game dialogue interfaces. In *Proceedings of the Fifth International Conference on the Foundations of Digital Games*, pages 179–186. ACM.

[Salvucci, 2001] Salvucci, D. (2001). An integrated model of eye movements and visual encoding. *Cognitive Systems Research*, 1(4):201–220.

[Salvucci and Taatgen, 2008] Salvucci, D. and Taatgen, N. (2008). Threaded cognition: an integrated theory of concurrent multitasking. *Psychological review*, 115(1):101.

[Salvucci and Taatgen, 2010] Salvucci, D. and Taatgen, N. (2010). *The multitasking mind*. Oxford University Press.

[Schatzmann et al., 2006] Schatzmann, J., Weilhammer, K., Stuttle, M., and Young, S. (2006). A survey of statistical user simulation techniques for reinforcement-learning of dialogue management strategies. *The knowledge engineering review*, 21(2):97–126.

[Schegloff, 1968] Schegloff, E. (1968). Sequencing in conversational openings. *American Anthropologist*, 70:1075–1095.

[Schlangen and Skantze, 2009] Schlangen, D. and Skantze, G. (2009). A general, abstract model of incremental dialogue processing. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 710–718. Association for Computational Linguistics.

[Schneider et al., 2015a] Schneider, J., Börner, D., Van Rosmalen, P., and Specht, M. (2015a). Presentation trainer, your public speaking multimodal coach. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 539–546, Seattle, WA, USA. ACM.

[Schneider et al., 2015b] Schneider, J., Börner, D., Van Rosmalen, P., and Specht, M. (2015b). Stand tall and raise your voice! a study on the presentation trainer. In *Design for teaching and learning in a networked world*, pages 311–324. Springer.

[Schön, 1983] Schön, D. (1983). The reflective practitioner: How professionals think in action. In Smith, T., editor, *Basic Books*. Temple Smith, London.

[Seabra Lopes and Teixeira, 2000] Seabra Lopes, L. and Teixeira, A. J. S. (2000). Human-robot interaction through spoken language dialogue. In *Proceedings IEEE/RSJ International Conf. on Intelligent Robots and Systems*, Japan.

[Searle, 1969] Searle, J. (1969). *Speech acts*. Cambridge University Press, Cambridge.

[Sebenius, 2007] Sebenius, J. (2007). Negotiation analysis: Between decisions and games. *Advances in Decision Analysis: From Foundations to Applications*, page 469.

[Seneff et al., 1998] Seneff, S., Hurley, E., Lau, R., Pao, C., Schmid, P., and Zue, V. (1998). Galaxy-II: A reference architecture for conversational system development. In *Proceedings of the 5th International Conference on Spoken Language Processing*.

[Serban et al., 2016] Serban, I., Sordoni, A., Bengio, Y., Courville, A., and Pineau, J. (2016). Building end-to-end dialogue systems using generative Hierarchical Neural Network models. In *AAAI*, volume 16, pages 3776–3784.

[Shang et al., 2015] Shang, L., Lu, Z., and Li, H. (2015). Neural responding machine for short-text conversation. *arXiv preprint arXiv:1503.02364*.

[Shotton et al., 2011] Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., and Blake, A. (2011). Real-time human pose recognition in parts from single depth images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[Sidner and Israel, 1981] Sidner, C. and Israel, D. (1981). Recognizing intended meaning and speakers' plans. In *IJCAI*, pages 203–208.

[Siegler and Stern, 1998] Siegler, R. and Stern, E. (1998). Conscious and unconscious strategy discoveries: A microgenetic analysis. *Journal of Experimental Psychology: General*, 127(4):377.

[Simpson, 1994] Simpson, G. (1994). Context and the processing of ambiguous words. *Handbook of psycholinguistics*, 22:359–374.

[Singh et al., 2017] Singh, M., Oualil, Y., and Klakow, D. (2017). Approximated and domain-adapted lstm language models for first-pass decoding in speech recognition. In *Proceedings of the 18th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Stockholm, Sweden.

[Singh et al., 2002] Singh, S., Litman, D., Kearns, M., and Walker, M. (2002). Optimizing dialogue management with reinforcement learning: Experiments with the NJFun system. *Journal of Artificial Intelligence Research*, 16:105–133.

[Smith et al., 1982] Smith, D., Pruitt, D., and Carnevale, P. (1982). Matching and mismatching: The effect of own limit, other's toughness, and time pressure on concession rate in negotiation. *Journal of Personality and Social Psychology*, 42(5):876.

[Sobol-Shikler, 2011] Sobol-Shikler, T. (2011). Automatic inference of complex affective states. *Computer Speech & Language*, 25(1):45–62.

[Sperry, 1993] Sperry, R. (1993). The impact and promise of the cognitive revolution. *American Psychologist*, 48(8):878.

[Spinuzzi, 2005] Spinuzzi, C. (2005). The methodology of participatory design. *Technical communication*, 52(2):163–174.

[Steiner et al., 2009] Steiner, C., Kickmeier-Rust, M., Mattheiss, E., and Albert, D. (2009). Undercover: Non-invasive, adaptive interventions in educational games. In *Proceedings of 80Days 1st International Open Workshop on Intelligent Personalisation and Adaptation in Digital Educational Games*, pages 55–65.

[Stevens et al., 2016a] Stevens, C., de Weerd, H., Cnossen, F., and Taatgen, N. (2016a). A metacognitive agent for training negotiation skills. In *Proceedings of the 14th International Conference on Cognitive Modeling (ICCM 2016)*.

[Stevens et al., 2016b] Stevens, C., Taatgen, N., and Cnossen, F. (2016b). Instance-based models of metacognition in the Prisoner's Dilemma. *Topics in cognitive science*, 8(1):322–334.

[Strangert, 1991] Strangert, E. (1991). Phonetic characteristics of professional news reading. *PERILUS*, XII:39–42.

[Strangert, 2005] Strangert, E. (2005). Prosody in public speech: analyses of a news announcement and a political interview. In *Proceedings of the 6th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 3401–3404, Lisbon, Portugal. International Speech Communication Association (ISCA), Baixas, France.

[Strangert and Deschamps, 2006] Strangert, E. and Deschamps, T. (2006). The prosody of public speech - a description of a project. *Lund Unversity Working Papers*, 52:121–124.

[Straßmann et al., 2016] Straßmann, C., von der Pütten, A., Yaghoubzadeh, R., Kaminski, R., and Krämer, N. (2016). The effect of an intelligent virtual agents nonverbal behavior with regard to dominance and cooperativity. In *International Conference on Intelligent Virtual Agents*, pages 15–28, Los Angeles, CA, US. Springer.

[Streitz, 2001] Streitz, N. (2001). Augmented reality and the disappearing computer. *Cognitive Engineering, Intelligent Agents and Virtual Reality*, 1:738–742.

[Su et al., 2016] Su, P.-H., Gasic, M., Mrksic, N., Rojas-Barahona, L., Ultes, S., Vandyke, D., Wen, T.-H., and Young, S. (2016). On-line active reward learning for policy optimisation in spoken dialogue systems. *arXiv preprint arXiv:1605.07669*.

[Sukhbaatar et al., 2015] Sukhbaatar, S., Weston, J., Fergus, R., et al. (2015). End-to-end memory networks. In *Advances in neural information processing systems*, pages 2440–2448.

[Sutskever et al., 2014] Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

[Sutton and Barto, 1998] Sutton, R. and Barto, A. (1998). *Reinforcement learning: An introduction*, volume 1(1). MIT press Cambridge.

[Swinney, 1979] Swinney, D. (1979). Lexical access during sentence comprehension: (re)consideration of context effects. *Journal of verbal learning and verbal behavior*, 18(6):645–659.

[Taatgen, 2013] Taatgen, N. (2013). The nature and transfer of cognitive skills. *Psychological review*, 120(3):439.

[Taatgen and Anderson, 2002] Taatgen, N. A. and Anderson, J. R. (2002). Why do children learn to say "broke"? A model of learning the past tense without feedback. *Cognition*, 86(2):123–155.

[Tanenhaus et al., 1995] Tanenhaus, M., Spivey-Knowlton, M., Eberhard, K., and Sedivy, J. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, pages 1632–1634.

[Tassinary, 2000] Tassinary, L. (2000). The skeletomotor system: Surface electromyography.

[Terken et al., 2006] Terken, J., van Dam, H., and Bunt, H. (2006). Cooperative assistance for human-system interaction. In Pikaar, R., Koningsveld, E., and Settels, P., editors, *Proceedings of the 16th World Congress on Ergonomics (IEA 2006, Congress of the International Ergonomics Association)*, pages 2004–2009. Elsevier.

[Tinsley et al., 2002] Tinsley, C., O'Connor, K., and Sullivan, B. (2002). Tough guys finish last: The perils of a distributive reputation. *Organizational Behavior and Human Decision Processes*, 88(2):621–642.

[Tomasello, 2009] Tomasello, M. (2009). *The cultural origins of human cognition*. Harvard university press.

[Touati, 1993] Touati, P. (1993). Prosodic aspects of political rhetoric. In *ESCA Workshop on Prosody*, pages 168–171, Lund, Sweden. International Speech Communication Association (ISCA), Baixas, France.

[Touati, 2009] Touati, P. (2009). Temporal profiles and tonal configurations in french political speech. *Working Papers in Linguistics*, 38:205–219.

[Toulmin, 1958] Toulmin, S. (1958). *The Uses of Arguments*. Cambridge University Press, Cambridge, England.

[Traum, 1994] Traum, D. (1994). *A Computational Theory of Grounding in Natural Language Conversation. PhD Thesis*. Department of Computer Science, University of Rochester.

[Traum, 1999] Traum, D. (1999). Computational models of grounding in collaborative systems. In Brennen, S., Giboin, A., and Traum, D., editors, *Working Papers of the AAAI Fall Symposium on Psychological Models of Communication in Collaborative Systems*, pages 124–131, Menlo Park, CA, USA. American Association for Artificial Intelligence.

[Traum, 2000] Traum, D. (2000). 20 questions on dialogue act taxonomies. *Journal of Semantics*, 17(1):7–30.

[Traum et al., 1999] Traum, D., Bos, J., Cooper, R., Larsson, S., Lewin, I., Matheson, C., and Poesio, M. (1999). A model of dialogue moves and information state revision. TRINDI project deliverable D2.1.

[Traum et al., 2008] Traum, D., Marsella, S., Gratch, J., Lee, J., and Hartholt, A. (2008). Multi-party, multi-issue, multi-strategy negotiation for multi-modal virtual agents. In *International Workshop on Intelligent Virtual Agents*, pages 117–130. Springer.

[Traum, 1993] Traum, D. R. (1993). Mental state in the TRAINS-92 dialogue manager. In *Working Notes AAAI Spring Symposium on Reasoning about Mental States: Formal Theories and Applications*, pages 143–149.

[Traum and Allen, 1992] Traum, D. R. and Allen, J. F. (1992). A "speech acts" approach to grounding in conversation. In *ICSLP*.

[Trouvain and Schröder, 2004] Trouvain, J. and Schröder, M. (2004). How (not) to add laughter to synthetic speech. *Affective Dialogue Systems*, pages 229–232.

[Tuppen, 1974] Tuppen, C. (1974). Dimensions of communicator credibility: An oblique solution. *Speech Monographs*, 41:3:253–260.

[Turunen et al., 2005] Turunen, M., Hakulinen, J., Raiha, K.-J., Salonen, E.-P., Kainulainen, A., and Prusi, P. (2005). An architecture and applications for speech-based accessibility systems. *IBM Systems Journal*, 44(3):485–504.

[van Dam, 2006] van Dam, H. (2006). *Dialogue acts in GUIs. PhD dissertation.* PhD thesis, Technische Universiteit Eindhoven, The Netherlands.

[van den Bosch and Bouma, 2011] van den Bosch, A. and Bouma, G. (2011). *Interactive multi-modal question-answering*. Springer Science & Business Media.

[Van Helvert et al., 2016] Van Helvert, J., Petukhova, V., Stevens, C., de Weerd, H., Börner, D., Van Rosmalen, P., Alexandersson, J., and Taatgen, N. (2016). Observing, coaching and reflecting: Metalogue - a multi-modal tutoring system with metacognitive abilities. *EAI Endorsed Transactions on Future Intelligent Educational Environments*, 16(6).

[Van Lehn, 2006] Van Lehn, K. (2006). The behavior of tutoring systems. *International journal of artificial intelligence in education*, 16(3):227–265.

[Van Rij et al., 2010] Van Rij, J., Van Rijn, H., and Hendriks, P. (2010). Cognitive architectures and language acquisition: A case study in pronoun comprehension. *Journal of Child Language*, 37(3):731–766.

[Van Rosmalen et al., 2015] Van Rosmalen, P., Börner, D., Schneider, J., Petukhova, V., and Van Helvert, J. (2015). Feedback design in multimodal dialogue systems. In Helfert, M., Restivo, M. T., Zvacek, S., and Uhomoibhi, J., editors, *Proceedings of the 7th International Conference on Computer Supported Education*, pages 209–217, Lisbon, Portugal. SCITEPRESS.

[Van Zanten, 1996] Van Zanten, G. V. (1996). Pragmatic interpretation and dialogue management in spoken-language systems. In *LuperFoy, S., Nijholt, A. and Veldhuijzen van Zanten, GE, TWLT11: Dialogue Management in Natural Language Systems, Proceedings of the Twente Workshop on Language Technology*, volume 11. Citeseer.

[Vapnik, 2013] Vapnik, V. (2013). *The nature of statistical learning theory*. Springer Science & Business Media.

[Veksler et al., 2012] Veksler, V., Myers, C., and Gluck, K. (2012). An integrated model of associative and reinforcement learning. Technical report, Air Force Research Lab Wright-Patterson AFB OH.

[Vinciarelli et al., 2009] Vinciarelli, A., Pantic, M., and Bourlard, H. (2009). Social signal processing: Survey of an emerging domain. *Image and vision computing*, 27(12):1743–1759.

[Wahlster, 2000] Wahlster, W., editor (2000). *Verbmobil: foundation of speech-to-speech translation*. Springer, Berlin, Heidelberg, New York, Barcelona, Hong Kong, London, Milan, Paris, Singapore, Tokyo.

[Walker et al., 1998] Walker, M., Fromer, J., and Narayanan, S. (1998). Learning optimal dialogue strategies: A case study of a spoken dialogue agent for email. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 2*, pages 1345–1351. Association for Computational Linguistics.

[Walker et al., 2000] Walker, M., Kamm, C., and Litman, D. (2000). Towards developing general models of usability with PARADISE. *Natural Language Engineering*, 6(3-4):363–377.

[Walker et al., 1997] Walker, M., Litman, D., Kamm, C., and Abella, A. (1997). PARADISE: A framework for evaluating spoken dialogue agents. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pages 271–280. Association for Computational Linguistics.

[Walton, 1996] Walton, D. N. (1996). *Argumentation schemes for presumptive reasoning*. Routledge, Oxford, UK.

[Walton and McKersie, 1965] Walton, R. and McKersie, R. (1965). *A behavioral theory of labor negotiations: An analysis of a social interaction system*. Cornell University Press.

[Wang et al., 2008] Wang, N., Johnson, W., Mayer, R., Rizzo, P., Shaw, E., and Collins, H. (2008). The politeness effect: Pedagogical agents and learning outcomes. *International Journal of Human-Computer Studies*, 66(2):98–112.

[Warner and Hirschberg, 2012] Warner, W. and Hirschberg, J. (2012). Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*, pages 19–26, Montreal, Canada. Association for Computational Linguistics.

[Watkins, 2003] Watkins, M. (2003). Analysing complex negotiations. *Harvard Business Review, December*.

[Watson et al., 2008] Watson, D., Tanenhaus, M., and Gunlogson, C. (2008). Interpreting pitch accents in online comprehension: H* vs. L+ H. *Cognitive Science*, 32(7):1232–1244.

[Welford, 1968] Welford, A. T. (1968). *Fundamentals of skill*. Methuen.

[Wen et al., 2017] Wen, T.-H., Vandyke, D., Mrksic, N., Gasic, M., Rojas-Barahona, L., Su, P.-H., Ultes, S., and Young, S. (2017). A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th European Chapter of the Association for Computational Linguistics (EACL 2017)*, Valencia, Spain.

[Whissell, 2009] Whissell, C. (2009). Using the revised dictionary of affect in language to quantify the emotional undertones of samples of natural language. *Psychological reports*, 105(2):509–521.

[Wichmann, 2002] Wichmann, A. (2002). Attitudinal intonation and the inferential process. In *Speech Prosody 2002, International Conference*, pages 11–16, Aix-en-Provence, France. Laboratoire Parole et Langage.

[Wilks and Ballim, 1991] Wilks, Y. and Ballim, A. (1991). Beliefs, stereotypes and dynamic agent modeling. In *User Modeling and User-Adapted Interaction*, volume 1(1), pages 33–65. Kluwer Academic Publishers, Dordrecht.

[Williams et al., 2013] Williams, J., Raux, A., Ramachandran, D., and Black, A. (2013). The dialog state tracking challenge. In *SIGDIAL Conference*, pages 404–413.

[Williams and Young, 2007] Williams, J. and Young, S. (2007). Partially Observable Markov Decision Processes for spoken dialog systems. *Computer Speech & Language*, 21(2):393–422.

[Woods et al., 2012] Woods, B., Aguirre, E., Spector, A., and Orrell, M. (2012). Cognitive stimulation to improve cognitive functioning in people with dementia. *Cochrane Database Syst Rev*, 2(2).

[Xu and Rudnicky, 2000] Xu, W. and Rudnicky, A. (2000). Task-based dialog management using an agenda. In *NLP/NAACL 2000 Workshop on Conversational Systems*.

[Young, 2000] Young, S. (2000). Probabilistic methods in spoken–dialogue systems. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 358(1769):1389–1402.

[Young et al., 2013] Young, S., Gašić, M., Thomson, B., and Williams, J. D. (2013). POMDP-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179.

[Yu et al., 2011] Yu, H.-F., Huang, F.-L., and Lin, C.-J. (2011). Dual coordinate descent methods for logistic regression and maximum entropy models. *Machine Learning*, 85(1-2):41–75.