ulm university universität
uulm

Faculty of Engineering, Computer Science and Psychology
Institute of
Communications Engineering

Multimodal Adaptive Dialogue Management in OwlSpeak

Master's Thesis at Ulm University

Submitted by:
Louisa Pragst

Examiner:
Prof. Dr. Dr.-Ing Wolfgang Minker
Prof. Dr. Enrico Rukzio

Supervisor:
Dipl. Inform. Stefan Ultes

2015

Issued: September 28, 2015

Satz: PDF-LaTeX 2$_\varepsilon$

Abstract

Spoken dialogue systems are employed in human-computer interaction to support the natural communication method of humans. Multimodal dialogue systems can provide an even more human-like interaction for the user by utilising additional communication channels besides speech, such as gestures, gaze, or facial expression. The ability to adapt the dialogue to the task, the situation, or the user, can further improve the user experience. This work focuses on the challenges of multimodal, adaptive dialogue systems for dialogue management.

A dialogue manager is the component of a dialogue system that chooses the next system action in dependence on the user action and the dialogue history. Additional input can be provided to the dialogue manager and influence its decision, thus enabling adaptation.

The work at hand examines models of emotion as well as culture, and the potential of their employment in dialogue management. Following these considerations, it describes adjustments of the OwlSpeak dialogue manager that enable it to handle the contemplated adaptations.

Moreover, this work illustrates the integration of the speech-based OwlSpeak dialogue manager into a multimodal dialogue system.

Thereafter, a user study is presented that determines the impact of the implemented adaptations discussed earlier in this work on the perceived naturalness of human-computer interaction.

In conclusion, the insights of this work are summarised and areas of future work proposed.

Contents

List of Figures

List of Tables

List of Dialogues

1 Introduction

Humans communicate with each other to exchange information and form strong social bonds, using language as well as facial expression, body posture, and many other indicators. From an early age, humans learn how to convey the intended semantic meaning. In this, they are influenced to a degree by the culture they grow up in, as the environment provides examples for humans to learn from. The importance of being able to communicate efficiently becomes obvious when considering the amount of — and ongoing interest in — writing and rhetorical training courses concerned with communication. Intercultural communication in particular needs special training in order to prevent misunderstandings.

When computers first emerged, punch cards were utilised for human-computer interaction. As this interaction is difficult and unintuitive for humans, soon researchers began to investigate speech as potential interaction modality, with the ultimate goal of providing a natural, human-like interaction. Over time, dialogue systems improved in many aspects, allowing for more complex dialogues, as well as utilising improved speech recognition and generation. Still, to achieve truly human-like communication, even more effort is necessary. In human-human communication, speech is not the only way to convey meaning. Multimodal dialogue system are able to improve the user experience by taking into account additional communication channels such as facial expressions. Also, the cultural background and emotional state of the user have a major impact in human-human communication, and dialogue systems that are adaptable to these factors may entail a more natural dialogue in human-computer interaction. The work at hand contributes to the aspiration of natural human-computer communication by enhancing the OwlSpeak dialogue manager [16] regarding adaptivity and multimodality.

The work done for this master's thesis, while being of general importance, was initiated in order to prepare OwlSpeak for use in the KRISTINA project. As this influences some decisions in the implementation process, a short overview of the project is given. This is followed by an outline of the work at hand.

1.1 The KRISTINA project

The KRISTINA project is funded by the *European Union's Horizon 2020 research and innovation programme* under grant agreement No. 645012. The project goal is to help immigrants in Europe with healthcare-related question by the means of a virtual agent. A multimodal dialogue system will be created for that purpose in the scope of the KRISTINA project.

The project incorporates a number of different use cases, one of which is partly implemented in the scope of this work.

The first use case addresses Turkish immigrants in Germany that have to take care of their elderly relatives. Like most people that care for their relatives, they often have no healthcare training and only a

limited amount of time, as they have jobs and other responsibilities they need to attend to. However, cultural misunderstandings and lack of knowledge about the German healthcare system additionally complicate the situation. For the caregiver, the KRISTINA agent can provide useful information concerning the proper care of the relative and the German healthcare system. Furthermore, it can suggest ways to deal with the additional pressure caused by taking care of their relative. The caretaker can be assisted by the KRISTINA agent in handling daily life. The agent can also act as a conversation partner in a limited amount of domains, to avoid the feeling of loneliness.
In a similar way, the KRISTINA agent can support Turkish elderly living in assisted-living facilities. The nurses of such facilities can benefit from the agent's cultural awareness.
Another target group of the KRISTINA project are Polish immigrants, working as caregivers for elderly in Germany. Similarly to the Turkish immigrants, they often lack professional training for this task, have limited knowledge of the German healthcare system and additionally might face intercultural challenges. The help provided to them by the KRISTINA agent resembles that provided for Turkish immigrants.
The last use case concerns Arab immigrants in Spain. They can request information regarding the Spanish healthcare system and general health advice.
As the medical domain that the KRISTINA agent will be employed in contains sensitive topics, it is especially important for the agent to be a trusted communication partner. It is therefore designed as socially competent, culturally aware, and emotion-sensitive. Dialogue management as the central component of a dialogue system is responsible for ensuring those qualities.

1.2 Outline

This work presents the steps taken to enable multimodal and adaptive dialogue management in OwlSpeak. It puts forward related work in Chapter 2, before giving a short introduction to dialogue management and OwlSpeak in Chapter 3.
Chapters 4 and 5 establish the theoretical basis for adaptability and multimodality in OwlSpeak, respectively. In accordance with the goals of the KRISTINA project, the emotional state and cultural background of the user is taken into account for the adaptation of the dialogue strategy. Existing models of culture and emotion are evaluated and adjusted to fit the goals of the work at hand. The requirements of multimodal dialogue management are identified by comparing a spoken and a multimodal dialogue system.
Chapter 6 describes the integration of OwlSpeak with the Visual SceneMaker authoring tool [12], resulting in a fully functional multimodal, adaptive dialogue system. A dialogue is modelled in accordance with the use cases of the KRISTINA project, enhanced by adaptability to the user's involvement in the conversation. The insights of Chapter 4 serve as a foundation for the adaptations implemented in the dialogue. The chapter is closed with the functional testing of the devised dialogue system.
Chapter 7 describes a user study conducted to ensure the effectiveness and suitability of the adaptations modelled in the dialogue. The results of this user study, as well as the implications for future research are presented.
The final chapter summarises the findings of this work and provides suggestions for future work.

2 Related Work

Multimodal as well as adaptive dialogue systems have received a lot of attention in recent years. While a full overview over the developed systems goes beyond the scope of the work at hand, some examples of related work are presented in this chapter.

This work focuses on adaptivity in regard to the cultural background and emotional state of the user. Both features were employed in several dialogue systems.

Pittermann et al. [30] create an emotion adaptive dialogue manager using a data-oriented approach. They employ a discrete model to represent the emotional state of the user, in combination with a statistical model for the policy.

The NIMITEK system [13] supports users while they solve problems such as the Tower-of-Hanoi puzzle in a graphic system. It employs a simple discrete model of emotion, differentiating between negative, positive and neutral emotional states. The dialogue manager chooses the time and the kind of support provided to the user, taking into account these states.

André et al. [1] present an emotion adaptive spoken dialogue system utilising a dimensional model of emotion. In response to the emotional state, stylistic variations related to politeness strategies are employed. This approach to adaptability is similar to the one proposed in the work at hand, however, a broader range of stylistic variations that are related to culture is considered in my approach.

Adaptation of the system's behaviour to cultural background is considered by Jan et al. [19]. However, the focus of their work is on aspects which are not relevant to dialogue management, such as proxemics, gaze, and overlap in turn taking.

Mascarenhas et al. [22] proposed a culturally adaptable model based on social status. The relationship between the system and user determines, how willing or reluctant the system is to interact with the user. This model has been employed in an application for inter-cultural training.

Adaptation to both emotion and culture has been realised in FAtiMA-PSI by Aylett [4] with the goal to increase cultural awareness in the user. However, it does not incorporate existing cultures, but relies on fictional alien cultures.

One of the earliest works on multimodal interaction is presented by Bolt [5]. He developed a system that can be operated by speech and pointing, whereby references like "that one" in combination with a pointing gesture can be resolved by the system.

In the scope of the COMIC project, Catizone et al. [7] explore multimodal dialogue management in internet applications. They consider modalities such as natural language, typed text, and pen input, as well as gestures, facial expression, and body posture.

McGlashan [23] revises dialogue management techniques developed for spoken dialogue systems. They are revised and extended for a multimodal system that combines spoken language with direct manipulation as input modalities.

3 Introduction to Dialogue Management and OwlSpeak

A basic understanding of dialogue management in general and the OwlSpeak dialogue manager in particular is necessary for the comprehension of the further work. Hence, a short overview of both topics is given in this chapter.

3.1 Dialogue Management

Dialogue management is an integral part of every dialogue system, as it determines the system's behaviour. The most important tasks of a dialogue manager are to keep track of the course of the dialogue and to decide on the next system action. Using the example given in Dialogue 3.1, the functionality of a dialogue manager is illustrated in the following.

The input of the dialogue manager is a semantic representation of the user input, called user move. In the example dialogue given, the semantic user move might be 'Request(Cure(Headache):?)'. The dialogue manager has to update the dialogue history to be able to access this information later. Then, it decides on the next system action. This can be done, for example, using static rules that were defined by the developer, or a statistical policy learned from real dialogue data. An adaptive dialogue manager takes into account additional factors for its decision, such as the emotional state of the user. The output given by the dialogue manager is a semantic representation of the system action, called system move. In the case of the example dialogue, a system move could take the form of 'Suggest(Action(User):Drink More)'. When the user rejects this solution with 'Reject(Action(User):Drink More)', the dialogue system needs to access the dialogue history to be able to recall the original task and offer a different solution.

The ability to perform additional tasks — besides the aforementioned — can improve the quality of the dialogue management. This includes, for example, grounding strategies to make sure the user's instructions have been correctly recognised and the approach to over-answering in case the user provides more information than currently expected by the system.

USER: How do I get rid of a headache?
SYSTEM: Perhaps you need to drink more.
USER: No, I drank a lot today.
SYSTEM: Aspirin can cure headaches.

Dialogue 3.1: In this example dialogue, the user consults the system because he has a headache.

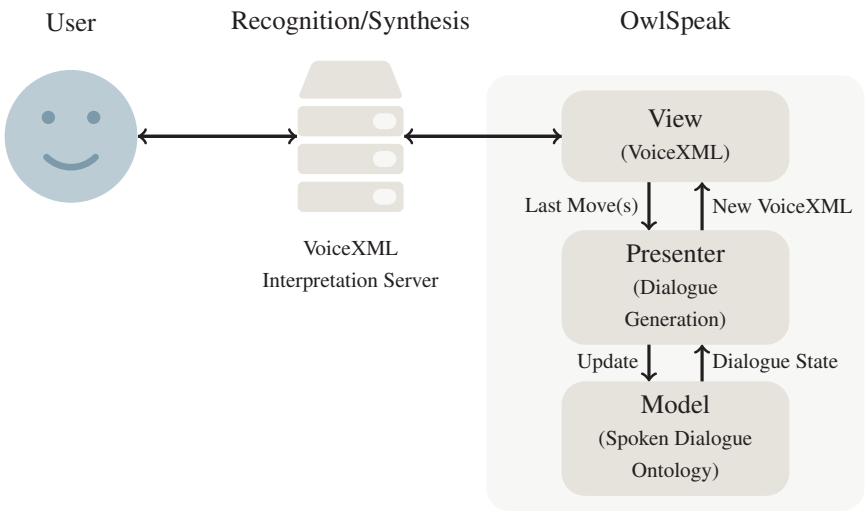

Figure 3.1: The general architecture of OwlSpeak (based on Heinroth et al. [16]) is consistent with the model-view-presenter pattern. The view in the form of VoiceXML is interpreted by a VoiceXML interpretation server. User input and output are speech based.

3.2 OwlSpeak

The OwlSpeak dialogue manager has been initially developed by Heinroth et al. [16] in the scope of the ATRACO project [14]. It is based on the information state theory by Larsson et al. [21] and is implemented in Java, following the model-view-presenter pattern [32]. This ensures the separation of data management, dialogue logic, and dialogue interface, which is beneficial for making modular changes as intended in this work. The general architecture can be found in Figure 3.1.
The model is embodied by a Spoken Dialogue Ontology (SDO) that is defined using the Web Ontology Language (OWL) [2]. A schematic representation can be found in Figure 3.2. The content of an SDO can be divided into static concepts representing the dialogue domain and dynamic concepts serving as the dialogue state. The concepts of SDOs are described in the following:

Grammar/Utterance *Grammars* are used in user moves to define what the user can say to actuate this specific move. In contrast, *utterances* are used in system moves and refer to one or more sentences that the system may say when performing this move.

Semantic/Variable Both concepts are used to represent information that is important for the dialogue, e.g. the meaning of what was said by the user. In contrast to *semantics*, *variables* can take on one out of several values and may be used for system-internal values.

Move A *move* can represent either a system or a user move, depending on whether it is related with a grammar or an utterance. A user move can be related to semantics that determine its meaning in the context of the dialogue.

Belief/BeliefSpace If a semantic is valid in the current dialogue state, this is represented by a corresponding *belief* being present in the *beliefspace*. Beliefs are set or unset (added to or removed from the beliefspace) by user moves that relate to the corresponding semantic.

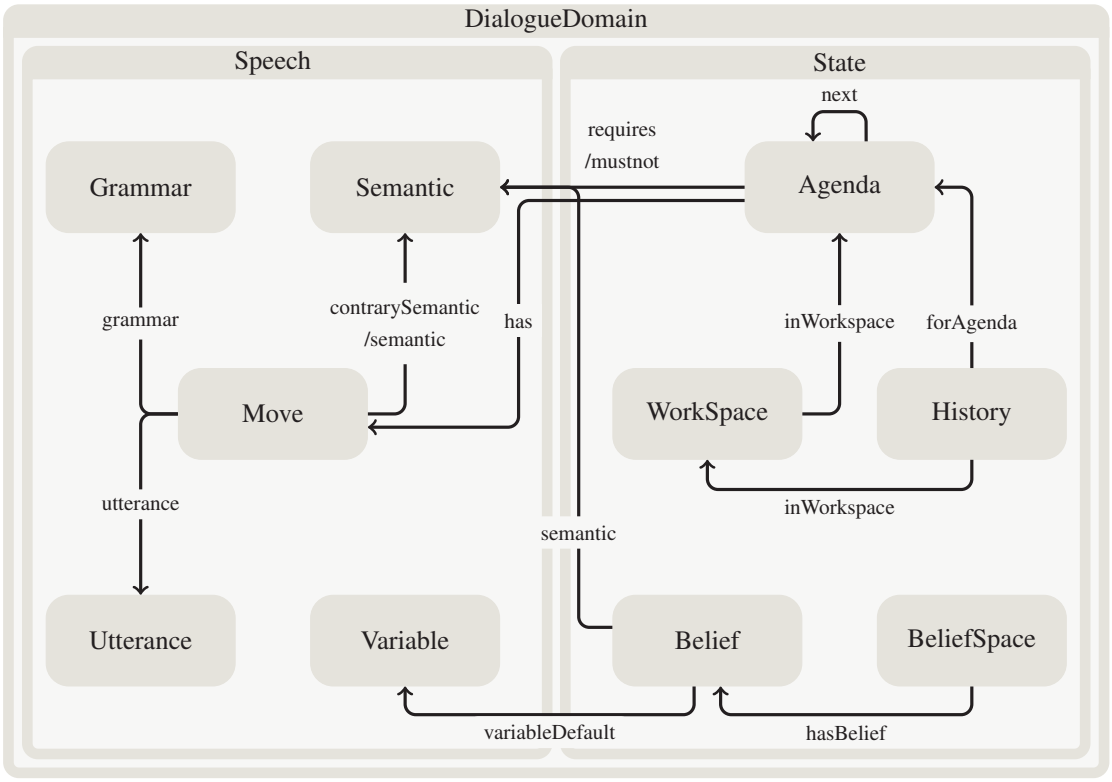

Figure 3.2: The Spoken Dialogue Ontology (as in Heinroth et al. [15]) is divided into a static dialogue description (Speech) and a dynamic dialogue description (State).

Agenda/WorkSpace An *agenda* relates to one system move as well as several user moves that may be expected from the user in response to the system move. Furthermore, it utilises the relation *next*, which denotes subsequent agendas that are added to the *workspace* upon the execution of the agenda itself. At each system turn, one agenda in the workspace may be chosen by the dialogue manager for execution. This choice is limited by the prerequisites of the agendas, denoted by *requires* and *mustNot*. For these prerequisites to be considered fulfilled, a corresponding belief has to be contained in or excluded from the beliefspace, respectively.

The first agenda of a dialogue is marked by a special *masteragenda* flag. Following its execution, the control of the dialogue is handled by the presenter component. It checks the prerequisites of all agendas in the workspace in order to determine the next system action. If more than one agenda is eligible for execution, the agenda with the highest priority score is chosen. This priority score can be predefined and/or dynamically assigned, depending on the policy, the amount of time the agenda is already in the workspace, or other factors.

The view represents OwlSpeak's interface, enabling the communication with the remaining components of the dialogue system. It is based on VoiceXML [28]; this is an XML document standard used to describe spoken dialogues. In combination with a VoiceXML interpretation server, OwlSpeak forms a functional spoken dialogue system. After each exchange between system and user, OwlSpeak generates a new VXML-document that is interpreted by the VoiceXML interpretation server to create audio output for the user.

The original OwlSpeak dialogue manager has been enhanced with additional functionality; Ultes et al. [38],

for example, applied the hidden information state approach [40] to OwlSpeak, creating HIS-OwlSpeak. Furthermore, adaptivity to Interaction Quality [34] has been implemented by Ultes et al. [37].

4 Towards Adaptation

Adaptive dialogue management makes dialogue systems more user friendly by modulating their behaviour in regard to the situation, the task, or the user. OwlSpeak already implements several adaptivity features: it is, for example, possible to adjust the dialogue strategy to the estimated interaction quality [37]. This work focuses on adaptation to the cultural background and emotional state of the user.
After establishing what to adapt to, two more issues have to be clarified. The first one is how to represent different manifestations of emotion and culture in the dialogue system. The second one is in what way the dialogue strategy should be modulated to match the different manifestations of culture and emotion. Existing models of emotion and culture are evaluated in the following, to find suitable representations and adaptations.

4.1 Adaptation to Culture

It has been observed (e.g. by Elliott et al. [9]) that different cultures prefer different communication styles. As a person is usually accustomed exclusively to the communication style of their own culture, it is reasonable to assume that talking to members of different cultures may lead to misunderstandings. The high amount of literature regarding business etiquette in foreign countries with the goal of reducing misunderstandings between business partners supports this claim.
By adjusting a system's behaviour to the culture of the user, misunderstandings can be reduced and the agent appears more familiar and therefore more trustworthy to the user.

4.1.1 Models of Culture

Communication Sciences are engaged in the classification of specific cultures in regard to their communication patterns. Models stemming from this research are particularly well suited for the use in adaptive dialogue management, as they provide dimensions along which cultures can be classified, and indicate how suitable adaptation strategies may be realised.
Prominent examples of cultural models include Hofstede's model of culture [17] or Kaplan's description of cultural thought patterns [20]. Hofstede's model of culture classifies cultures in regard to the values most frequently hold by its members. The dimensions are power distance, individualism, masculinity, uncertainty avoidance, long term orientation, and indulgence. These dimensions are suitable to influence the semantic content of system moves. Kaplan's cultural thought patterns are based on typical argumentative structures used by members of a culture, and differentiates between English, Semitic, Oriental, Romance, and Russian thought patterns.

[The system uses the linear argumentative style.]
SYSTEM: You should drink more. It is hot, therefore your body needs more water.

[The system uses the parallel argumentative style.]
SYSTEM: You should drink more. It is hot and your body needs more water.

[The system uses the circular argumentative style.]
SYSTEM: About 60% of the human body is water. Water is important for the proper functioning of your body.

[The system uses the digressive argumentative style.]
SYSTEM: You should drink more. Eating enough is also very important, but when it is so hot, your body needs more water.

Dialogue 4.1: Examples of different argumentative styles.

However, the dimensions used by Elliott et al. [9] to compare communication styles of different cultural groups in the United States prove to be better suited for dialogue management. They include the information of the aforementioned models in addition to other important aspects of communication. The work at hand disregards dimensions such as vocal pattern and gestures, as dialogue management is not responsible for decisions concerning the presentation of output. This leaves the dimensions that are considered relevant to dialogue management: animation/emotion, thought pattern/rhetorical style, directness/indirectness, identity orientation, turn taking/pause time, and time. Hofstede's dimensions are included in this list as animation/emotion, identity orientation and time (Hofstede's model of culture is more exhaustive in this regard). Kaplan's cultural thought patterns are represented as thought pattern/rhetorical style.
In the following, a short description of each dimension is presented:

Animation/Emotion The display of emotions and the apparent involvement in a topic can be perceived very differently across cultures. While in some cultures, strong emotions show that the user is adamant about their opinion, in others this behaviour may seem as uncontrolled, exaggerated, or aggressive.

In dialogue management, this influences the interpretation of the user's behaviour and also results in different system moves being chosen as response. While the dialogue manager will mostly stick to purely 'functional' dialogue moves in cultures with a restrained communication style, more expressive communication styles require dialogue moves expressing emotion, if applicable even without any functional, semantic content.

Thought Pattern/Rhetorical Style This aspect refers to the way arguments are presented in a discussion. The rhetorical styles of Kaplan are characterised by linearity, parallelism, circularity, and digression, respectively. Examples can be found in Dialogue 4.1.

When using a linear rhetorical style, arguments are presented sequentially and hierarchically sorted. Following the parallel style, however, the arguments do not have a hierarchy and are presented in a parallel manner, using coordinators, such as *and*, *but*, and *or*, rather than subordinators, such as *since*, *because*, and *although*. The circular style often does not elaborate on the main topic, but discusses a multitude of topics that might have an impact. Finally, the digressive style discusses not

[The system uses direct verbalisation]
SYSTEM: Take an aspirin.

[The system uses indrect verbalisation]
SYSTEM: Aspirin can help with headaches.

Dialogue 4.2: Examples of direct and indirect verbalisation of the speaker's intent.

only the main topic, but also related topics.

The rhetorical style that is to be realised has a great impact on the strategy pursued by the dialogue manager. An argument can be built over several dialogue moves, and needs to be consistent and follow one strategy in order to be convincing. Taking rhetorical style into account helps the dialogue management to provide the necessary information to the user more appropriately so that the user is more likely to accept it.

Directness/Indirectness While in some cultures it is favoured and expected to directly express your intent, others prefer a more indirect communication style whereby the listener has to deduce the intent from the context. Failure to do so can even be perceived as aggressive. Dialogue 4.2 presents a direct and an indirect way to propose aspirin as treatment for a headache.

Taking the directness level of a culture into account results in a suitable presentation of the information provided for the user. Depending on the specific architecture, this might be achieved either by different dialogue moves chosen by the dialogue manager or by speech generation expressing the same intent in different ways.

Identity Orientation Humans have internalised self-perception and certain values that influence their decisions. In some cultures these tend to be more group oriented: one's status is dependent on the status of one's family, and decisions are often made considering the well-being of the social group. In contrast to that there are also more individualistic cultures, in which one's status is dependent on one's own achievements. Decisions in such cultures tend to be made considering one's own well-being before that of the group.

Knowledge about the identity orientation can be used by the dialogue management in order to determine what to propose and which arguments to use to convince the dialogue partner. An example of this can be found in Dialogue 4.3.

Turn Taking/Pause Time There are many different ways to signal one's communication partner that it is their turn to speak. Often, pauses are used, in which case after some period of silence — that can greatly vary in length depending on the culture — the partner may speak. Upon speaking before enough time has passed, a person is often considered to not be thinking before talking. In contrast, in some cultures it is perfectly normal to interrupt each other in a conversation. Another option that is used in some cultures is eye contact, for example by looking at someone when they are supposed to speak.

Turn taking is relevant to the dialogue manager insofar as interruptions are concerned. Dialogue management has to be able to correctly interpret interruptions and handle them in a suitable way. The directing of gaze and maintaining of pauses is not part of dialogue management. Keeping pauses

[The system motivates the user using group oriented arguments.]
SYSTEM: You're a big help for your family.

[The system motivates the user using individualistically oriented arguments.]
SYSTEM: It is impressive that you are able to handle all of this.

[The system persuades the user using group oriented arguments.]
SYSTEM: This is the established way to lift a person.

[The system persuades the user using individualistically oriented arguments.]
SYSTEM: The way you lift the patient damages your back in the long run. This is a better way to do it.

Dialogue 4.3: Examples of arguments based on different values.

for the right amount of time might be a particularly hard challenge for the real-time requirement of any spoken dialogue system.

Time Punctuality does not have the same value in every culture. In some cultures, it is considered highly impolite to be late to an appointment, whereas in others waiting for the 'right' time is deemed proper and insistence on punctuality is rude behaviour.
While a dialogue system will not have problems with punctuality itself, this aspect might be considered when reminding users of appointments. Suitable system moves for every culture have to be implemented and chosen accordingly.

The dimensions chosen for this work are specifically concerned with the classification of cultures in regard to communication patterns. This facilitates their implementation in a dialogue manager. During the design of the dialogue, all targeted cultures need to be considered and incorporated as suitable system and user moves. The culture of the user can then be used to choose the appropriate system action.

4.1.2 Future Work

To utilise the presented model, it is necessary to map all cultures to the presented dimensions. A lot of work has been done in this regard already. However, the KRISTINA use cases include people with Turkish, Polish, German and Arab background, and the classification of these cultures is not sufficiently well researched. In the scope of the KRISTINA project, the classification of these cultures will have to be amended.
Of the considered cultures, Arab communication patterns have received the most attention in research and are described in detail, for example, by Feghali [10], Kaplan [20], and Zaharna [42]. However, their findings should not be adopted unrevised in KRISTINA for several reasons: first, the Arab culture is not well defined and often includes subcultures that could be regarded as autonomous cultures in their own right (the KRISTINA project focuses on the Moroccan subculture). Second, the cited papers date back more than fifteen years. As cultural preferences can change over time, the finding might no longer be valid. Third, the papers did not focus on the medical domain which is the centre of attention in the KRISTINA project. It is possible that the domain influences the communication patterns. For these reasons, it is

necessary to re-evaluate Arab communication patterns for the KRISTINA project.
Information about the remaining communication patterns can be found primarily in business guides and might not be applicable due to the different domain of the KRISTINA project.

4.2 Adaptation to Emotion

The medical domain poses a difficult area of conversation as often sensitive and personal topics need to be discussed. Several studies (e.g. Jaksic et al. [18], Partala et al. [29]) show that taking into account the emotional state can improve the user acceptance of the system.
In the following, different models of emotion are presented and considered regarding their applicability in dialogue management.

4.2.1 Models of Emotion

Many models of emotion have been proposed and employed in dialogue systems. This section provides an overview of the most frequently utilised models.
Plutchik [31] pursues a discrete approach to model emotions, called the wheel of emotions. It contains eight basic emotions: anticipation, anger, disgust, sadness, surprise, fear, trust and joy. In addition, these emotions can vary in their intensity, e.g. vigilance — anticipation — interest, and eight derivative emotions are defined, each of which is composed of two basic emotions, e.g. optimism is derived from anticipation and joy.
Another discrete model is commonly called the *Big Six*, and comprises anger, disgust, fear, happiness, sadness and surprise. These six emotions have been identified as basic emotions by Ekman et al. [8]. By conducting several studies, the Big Six have been shown to be recognisable across different cultures on the basis of facial expressions. It is therefore assumed that they are basic human emotions shared by all cultures.
The PAD (Pleasure — Arousal — Dominance) model [24] by Mehrabian describes emotional states using a three-dimensional, continuous space. It is therefore more flexible and precise than discrete models of emotion. Discrete emotions such as joy can be mapped into the PAD space.
In contrast to the presented cultural models, which originate from communication sciences and were designed to classify cultures in regard to conversational aspects, the emotional models have their roots in psychology and are not directly connected to communication. Hence, the mapping of the emotional state of the user to a suitable dialogue strategy is not as straightforward. While all mentioned models were employed in dialogue systems, often the adaptation of the dialogue strategy was limited to adjustments of the propositional contents. An exception is the work of André et al. [1]. A more general approach that enables a broad range of variations in rhetorical style, as well as semantic content, in response to the user's emotional state, would be desirable.
As described in the previous section, a deviation in rhetorical style from the cultural norm often carries emotional meaning — for example, directness can be perceived as aggression by members of more indirect cultures. Considering this observation, a new perspective presents itself. It can be reasonably assumed that

the presented cultural dimensions may be used as emotional dimensions as well. This would support the desired general adaptation strategies and facilitate the integration of the cultural and the emotional model. For the realisation of this concept, a mapping from classical models of emotion to these new dimensions is needed.

4.2.2 Future work

Although emotions are regularly used for adapting the behaviour of dialogue systems, a systematic mapping from emotional models to communication patterns has not yet been established.
Possibly, the dimensions of the cultural model are suitable to cater to the emotional state of the user. A preliminary user study has been conducted in the scope of this work, in order to determine the influence of rhetorical style on the perceived naturalness of a dialogue. The findings of this study are presented in Chapter 7.
Irrespective of these findings, further user studies are needed in order to determine suitable general dialogue strategies in regard to the user's emotional state.

4.3 Implications for Adaptation in OwlSpeak

Adaptability requires knowledge about the current state of the feature the dialogue is adapted to, as well as variants of system moves that can be chosen in accordance with that state.
In OwlSpeak, this can be achieved by adding variables representing the dimensions of the discussed models to the dialogue domain, e.g. *pleasure* from the PAD model. User moves set the value of those variables: 'I am fine.' results in a pleasure level of 1, 'I feel down.' in a pleasure level of -1.
In addition to knowing about the user's state, the system needs to react in different ways to different states. Therefore, the second step is to provide different variants of system moves. Prerequisites for these system moves, such as 'pleasure > 0.7' ensure that they are only chosen when appropriate.
Considering this, OwlSpeak already offers the mechanisms needed to enable adaptability. However, the quality of the adaptation depends on the amount of thought put into the design of the dialogue domain.
The steps taken in the scope of this work to enable adaptability are described in more detail in Chapter 6.

5 Towards Multimodality

After identifying the requirements for adaptivity in dialogue management in the previous chapter, this chapter is dedicated to the requirements imposed on dialogue management by multimodality. To this end, the functioning of a spoken dialogue system is compared to that of a multimodal dialogue system. Afterwards, requirements are deduced from the observed differences.

5.1 Comparison of a Spoken and a Multimodal Dialogue System

Figure 5.1 shows the general architecture of a spoken dialogue system. It incorporates speech recognition and semantic analysis in the input layer, as well as text generation and a text-to-speech component in the output layer. At the core, a dialogue manager updates the dialogue history, communicates with the underlying application and decides on the next system action. The interaction of these components is described in the following.

The speech recognition component gets audio input, e.g. the user saying 'How do I get rid of a headache?', and produces the content of the audio data in text. The linguistic analysis then interprets the semantic content of this text and forwards its findings in the form of a user move, such as 'Request(Cure(Headache):?)', to the dialogue manager. The dialogue manager could decide to ask further questions or get the needed information from the underlying application, and produce a semantic system move, for example 'Inform(Cure(Headache):Aspirin)', as output. This system move is then translated into a sentence by the text generation: 'Aspirin can cure headaches.'. Finally, the speech synthesis transforms this sentence into an audio output.

A multimodal dialogue system is characterised by the fact that more than one in- and output modality is utilised, in contrast to a spoken dialogue system, which employs only one modality: speech. Figure 5.2 shows the architecture of such a multimodal dialogue system, exemplary with the two modalities speech and gestures.

The input and output layers for speech do not differ from the corresponding layers of the spoken dialogue system. For gestures, components get included that perform analogous tasks: gesture recognition maps the initial input, e.g. video from a camera, to a set of predefined gestures such as 'Hand to Head'. Gesture interpretation might then conclude that in this situation this carries the semantic meaning 'Emphasise(Head)'. The gesture generation component gets the semantic input 'Show(Aspirin)' and translates that to the gestures 'Take Aspirin' and 'Hold Aspirin in Front of Body'. Finally, the agent animation produces a video of the agent performing this gesture.

Apart from those additional components that perform established tasks for a different modality, two more components are introduced: fusion and fission. The fusion module is responsible for synchro-

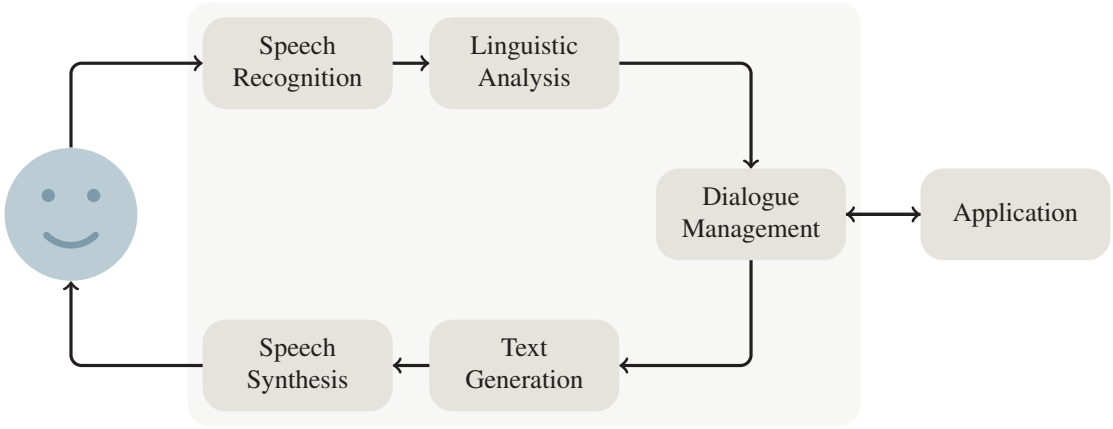

Figure 5.1: The general architecture of a spoken dialogue system (based on Minker [27]).

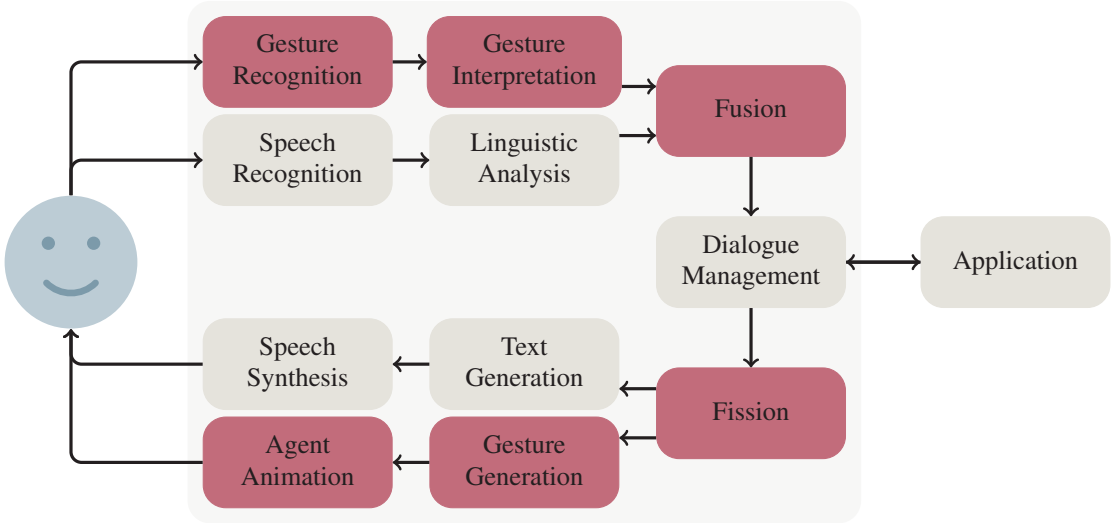

Figure 5.2: The general architecture of a multimodal dialogue system (based on Minker [27]). Fusion and fission components have to be added in order to handle the different modalities.

nising the incoming semantic representations and combine them into a single semantic user move. In the case of 'Request(Cure(Headache):?)' and 'Emphasise(Head)', this does not alter the user move 'Request(Cure(Headache):?)', but strengthens the confidence in the correctness of the interpretation. An example for additional information added by gestures would be pointing at someone while asking 'What's his name?'. In both cases, a single semantic user move is forwarded to the dialogue manager just as in a spoken dialogue system.

The output of the dialogue manager stays the same as well: a semantic system move is provided to the fission module. It is the output layer equivalent of the fusion component, splitting the system move to semantic representations for each modality. It is the responsibility of this component to decide which part of the system move 'Inform(Cure(Headache):Aspirin)' should be conveyed by which output modality. For instance, the text generation could receive the modified system move 'Inform(Cure(Headache):This)', and by adding the Gesture 'Show(Aspirin)' the whole message would be delivered to the user.

Of course, multimodal dialogue system may use more than two modalities. The architecture, however, stays almost identical: additional components have to be introduced to perform recognition, interpreta-

tion, generation, and output transformation, while the fusion and fission components take into account all of the utilised modalities. The dialogue manager continues to interact solely on a semantic level with the fusion and fission component, respectively.

5.2 Implications for Multimodality in OwlSpeak

By comparing the architecture of a spoken dialogue system to the architecture of a multimodal dialogue system, it becomes clear that the dialogue manager remains unaffected by multimodality to a great extend. The fusion module receives the semantic representations from all modalities and transforms them into a single semantic representation similar to the one the dialogue manager would receive in a spoken dialogue system. As output, the dialogue manager continues to provide a semantic system move, and the responsibility to distribute this move among all modalities lies with the fission module.
Therefore, in order to make OwlSpeak able to handle multimodality, two tasks need to be addressed: choosing an appropriate software component that can handle multimodal input and output, to replace the VoiceXML Interpretation Server, and adapting the view of OwlSpeak to enable the interaction with this new component. The following chapter describes the realisation of these tasks.

6 Realisation of Multimodal Adaptive Dialogue Management

In the scope of this work, the spoken dialogue manager OwlSpeak was enhanced regarding adaptivity and multimodality. This chapter presents the work done in this endeavour.
First, Visual SceneMaker [12] is introduced, the software component integrated with OwlSpeak to form a multimodal dialogue system. Following this, the changes implemented in Visual SceneMaker and OwlSpeak are described and the dialogue domain that was modelled for this thesis is presented. Finally, the functional testing of the devised dialogue system is outlined.

6.1 Visual SceneMaker

Visual SceneMaker [12] is a tool for creating interactive applications with virtual characters, which is implemented in Java. It has been chosen for the integration with OwlSpeak in this work because the source code is publicly available [11], which facilitates performing the necessary changes, and because Visual SceneMaker already utilises the two modalities audio and video as output, and supports the employment of several input modalities. Additionally, Visual SceneMaker offers the possibility to provide idle behaviour of the agent. In the following, a short overview of Visual SceneMaker is given in order to aid the understanding of the implementational issues of the work at hand.
The two most important concepts of Visual SceneMaker are *scenescripts* and *sceneflows*. A scenescript contains instructions for the virtual agent: mainly the intended phrases, but also directions on gestures. The content of the agent's utterance can be modulated by the use of variables. A sceneflow is a hierarchical statechart variant used to define the course of events in the application. It can be regarded as a simplified dialogue manager.
A sceneflow can be created with the graphical user interface of Visual SceneMaker. Its nodes may be associated with the playback of individual scenes, the execution of commands or the processing of user interactions. Different kinds of edges — such as conditional or probabilistic edges — connect the nodes to define the workflow of the application.
The output of Visual SceneMaker is generated from the scenescripts by the Horde3D graphics engine [35]. The Social Signal Interpretation (SSI) framework [39] is responsible for the processing of user input. SSI supports a broad range of sensor devices, filters and feature algorithms as well as machine learning and pattern recognition tools. It is therefore suitable to process a large variety of modalities as input. For speech recognition, SSI utilises the Microsoft Speech Platform SDK 11 [26].

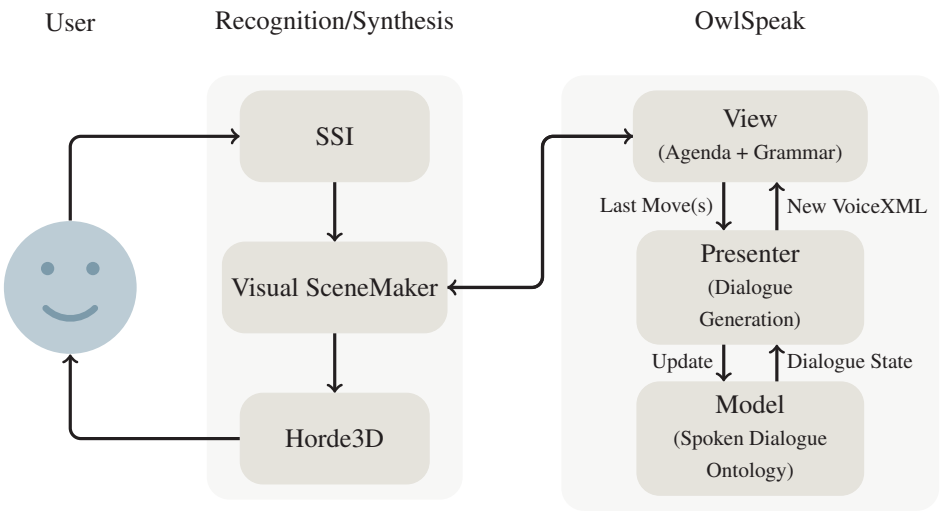

Figure 6.1: The general architecture of a multimodal dialogue system that utilises Visual SceneMaker and OwlSpeak. OwlSpeak's architecture is still consistent with the model-view-presenter pattern. The view has to be adapted in order to enable interaction with Visual SceneMaker.

6.2 Implementation

For this thesis, OwlSpeak has been integrated with Visual SceneMaker in order to render it capable of multimodality. The resulting architecture is shown in Figure 6.1. The following changes were performed to achieve multimodality: a new Java class called *OwlSpeakEngine* was added to Visual SceneMaker as interface connection to OwlSpeak, the view of OwlSpeak was modified, and a grammar parser was implemented. Furthermore, a sceneflow was created, an SDO (which is described in more detail in the next section) designed and scenescripts corresponding to the system moves of the SDO were written. In the following, the interaction of these components is described.

Upon the start of Visual SceneMaker, a method of *OwlSpeakEngine*, which starts OwlSpeak, is called. Afterwards, the execution of the sceneflow starts and the application waits for SSI to detect user input. When SSI passes the detected user input to Visual SceneMaker, the sceneflow arrives at a node that calls another method of *OwlSpeakEngine*. This method receives the detected user move and forwards it to OwlSpeak. OwlSpeak processes the user move in its regular way and provides the next agenda. The agenda is no longer presented as VXML-document. Instead, it consists of the identifier of the system move of the agenda, and a grammar. The identifier of the system move is used to determine the scenescript that is to be played back to the user. The next node in the sceneflow is responsible for providing this output, before the application returns to the start of the sceneflow and waits for further user input. The grammar is used by SSI for the speech recognition. OwlSpeak produces grammars by combining the individual grammars of each user move contained in the chosen agenda. Each possible user utterance is tagged with the user move it belongs to. This enables speech recognisers to identify which user move has been enacted. As SSI uses the Microsoft Speech API, the created grammar has to be parsed to the SAPI Grammar format by OwlSpeak when generating the view. A parser was developed to perform this task.

The described setup uses speech input only. In order to utilise the potential for multimodality, the setup of SSI was changed to detect one of Plutchik's [31] emotions: interest in the dialogue, or user involvement.

A lot of research has been conducted in regard to detecting involvement. Yu et al. [41] utilize Support Vector Machines and Hidden Markov Models to detect user involvement in natural speech, Asteriadis[3] consider head pose and eye gaze, Sanghvi [33] derive engagement from posture and body motion, while Szafir et al. [36] use EEG measurements. As emotion detection is not in the focus of this work, a rudimentary approach to detecting involvement has been chosen: user interest is determined by the amount of time the user's face is directed at the avatar, averaged over time. This is done using the face tracking feature of the Kinect [25], which was already possible in SSI. If the user turns his head away from the screen and is not paying attention to the avatar, the face can no longer be detected by the Kinect and the estimated user involvement decreases.
The additional information about the user's interest in the dialogue can be utilized by OwlSpeak when choosing the next agenda, rendering it adaptive to user involvement. This is achieved by adding the detected user interest as variable value to the detected user move before passing the user move to OwlSpeak. When processing this user move, OwlSpeak's beliefspace is updated accordingly and only agendas that are suitable for the level of user involvement can be chosen. The next section describes in detail the modelled dialogue domain and its adaptation potential.

6.3 Dialogue Design

Before a dialogue system can be employed, a suitable dialogue has to be designed. This requires the specification of the intended use case scenario and all anticipated user and system moves. Considerations regarding the intended influences and adaptation possibilities have an impact on this process.
For this work, the use case scenario is adopted from the KRISTINA use cases. More than one use case of the KRISTINA project is concerned with basic care-giving advice for untrained personnel. The modelled dialogue reflects this problem by providing tips regarding fluid intake and headaches. This is a basic example that has been regularly mentioned by the use case partners during meetings. A full overview of the modelled dialogue can be found in Figures 6.2-6.14. In the following, the possibilities for adaptation integrated in this dialogue are described.

6.3.1 Design Decisions regarding Adaptation

The modelled dialogue needs to offer possibilities for adaptation in order to enable adaptability of the implemented dialogue system. In Chapter 4, it was proposed that the variations in rhetorical style used to classify cultures may be suitable to react to emotions of the user. The dialogue modelled for this work realises such variations of rhetorical style in response to the level of user involvement.
The task described in the scenario is information retrieval. The user asks for some information; the system may ask for further information which it needs in order to be able to answer the question, or provide the answer directly. Often, such systems are evaluated by the number of moves needed to answer the question. The preferred strategy is the one that can provide the information the user needs as fast as possible. This implicitly results in choices for rhetorical style, such as directness, concision and a linear course of conversation. An example of such a conversation is presented in Dialogue 6.1.

SYSTEM: Hello <User>.
USER: [*uninvolved*] Hello Kristina.
SYSTEM: How can I help you?
USER: [*uninvolved*] I'm afraid Mr. Smith doesn't drink enough during the day.
SYSTEM: How much does Mr. Smith usually drink?
USER: [*uninvolved*] Only one cup of tea after breakfast.
SYSTEM: You're right, that is not enough. Do you know why he doesn't drink enough?
USER: [*uninvolved*] He doesn't like the taste of water. He would rather have some soft drink, but I don't think that is good for him.
SYSTEM: Then you should try sweetened tea. It tastes good and is not as bad as a soft drink.
USER: [*uninvolved*] Perhaps I don't even need to sweeten it. He does drink tea in the morning, after all.
SYSTEM: Yes, you can try that.
USER: [*uninvolved*] And if that doesn't work. Any other suggestions?
SYSTEM: I have got three more ideas. Do you want to hear all of them?
USER: [*uninvolved*] Perhaps two more. You can tell me the last one another time.
SYSTEM: First, you should keep him company. While chatting people usually drink a lot without even realising it. Second, you can increase his fluid intake by preparing a soup for him as it contains a lot of fluid.
USER: [*uninvolved*] Alright. Thank you, Kristina. You really helped me.
SYSTEM: No problem at all.
USER: [*uninvolved*] I'll get back to work.
SYSTEM: I will be here if you need me.

Dialogue 6.1: Example dialogue presenting a dialogue system that utilises a direct, concise and linear rhetorical style.

While this strategy is often a good choice, especially if the user is in a hurry or distracted, and has no interest in a lengthy conversation, it may be unsuitable under specific circumstances. Examples for such circumstances can be the cultural background or emotional state of the user, as discussed in Chapter 4.

For an uninvolved user, the described dialogue strategy is suitable and corresponds to the behaviour in natural human-human communication. However, the intended use of the KRISTINA agent goes beyond information retrieval. The agent is supposed to be a long-term companion of the user and form a social bond with them. In such a scenario, a high user involvement can be expected and should be supported.

An example dialogue of what I believe to be a suitable dialogue strategy for high user involvement is given in Dialogue 6.2. It adheres to the following rhetorical choices:

First, social conventions can be taken into account when choosing system moves, even though such moves do not contribute to the content of the dialogue. An example for this can be seen in Line 3 of Dialogue 6.2; the KRISTINA agent asks the user: 'How are you?'. This would be inconvenient for an uninvolved user that just wants to retrieve the needed information as quickly as possible. However, for an involved user, this makes the conversation more human-like and supports social bonding.

Second, answers can be more verbose. While this consumes additional time and can be distracting from the important information, and therefore is not desirable for an uninvolved user, for the involved user, this makes the dialogue more animated and can offer links for further conversation topics. It can even be considered to include information that was not requested by the user, but is related to the topic. An example for this can be found in Line 9 of the dialogue.

SYSTEM: Hello <User>.
USER: [*involved*] Hello Kristina.
SYSTEM: How are you today?
USER: [*involved*] I'm fine, thank you.
SYSTEM: Great. And how can I help you?
USER: [*involved*] I'm afraid Mr. Smith doesn't drink enough during the day.
SYSTEM: I see. How much does Mr. Smith usually drink?
USER: [*involved*] Only one cup of tea after breakfast.
SYSTEM: You're right, that is not enough. In these temperatures, Mr. Smith should drink at least 1.5 litres. Do you know why he doesn't drink enough?
USER: [*involved*] He doesn't like the taste of water. He would rather have some soft drink, but I don't think that is good for him.
SYSTEM: You are right, water is the better choice. Sweetened tea may be a compromise. It has a better taste than water and less sugar than a soft drink.
USER: [*involved*] Perhaps I don't even need to sweeten it. He does drink tea in the morning, after all.
SYSTEM: Exactly. A little company may help as well. While sitting and chatting you can easily drink some glasses without even noticing it.
USER: [*involved*] I guess that's true. Chatting is just more comfortable with a nice cup of tea.
SYSTEM: By the way, does Mr. Smith use a clear glass?
USER: [*involved*] Yes, he does. A beer glass with a handle, so he can better grab it. Why?
SYSTEM: That might be too unobtrusive. A clear glass filled with water does not really attract attention.
USER: [*involved*] Do you think something more eye-catching would be better.
SYSTEM: It is worth a try. Perhaps something more colourful?
USER: [*involved*] I could use the cup his granddaughter gave him. It's very colourful.
SYSTEM: That is a very good idea. He will be happy whenever he sees that cup.
USER: [*involved*] I think so, too. He loves her very much. Also, that cup has a handle. That is very important for him, he feels more secure when grabbing something with a handle.
SYSTEM: Many elderly feel that way.
USER: [*involved*] But what if Mr. Smith still doesn't drink enough? After all, I need to increase that quite a lot.
SYSTEM: The body doesn't only get fluid by drinking. Food contains fluid too.
USER: [*involved*] I don't think that will be enough.
SYSTEM: There some dishes that contain a lot of fluid, such as soup.
USER: [*involved*] Or goulash? He really likes that.
SYSTEM: Yes, goulash is fine. And fruits.
USER: [*involved*] Alright. Thank you, Kristina. You really helped me.
SYSTEM: No problem at all.
USER: [*involved*] I'll get back to work.
SYSTEM: I will be here if you need me.

Dialogue 6.2: Example dialogue presenting a dialogue system that utilises an indirect, verbose and spontaneous rhetorical style.

Third, the course of the dialogue does not have to be as linear as in a dialogue with low user involvement. A linear, structured course of dialogue imposes less cognitive load on the user, however, it also appears to be planned in advance. In human-human communication, this is plausible only if the conversation partner is a professional, whose only task is to answer questions of the kind asked. It is unlikely in a spontaneous dialogue and can therefore be perceived as less natural. Hence, an unstructured course of dialogue could contribute to the perceived naturalness. The modelled dialogue realises this, for example in Line 19, by introducing related topics in a spontaneous way, indicated by phrases like 'by the way'.
Fourth, indirect verbalisation of proposals can be used. In contrast to a direct verbalisation, this results in a higher cognitive load, as the message has to be inferred. The advantage is that it ensures the continued attention of the listener and offers more possibilities for participation in the dialogue. The message can be rephrased, follow-up question are more likely and one's own conclusions can be proposed. Also, influencing the topic of the dialogue is easier. An example of this can be found in Line 21 of the dialogue.
Finally, the amount of information given can be varied. While incomplete information in areas the user cannot have knowledge of is not recommendable, omitting information that the user might have gives him the opportunity to contribute this knowledge to the conversation themselves, as can be seen in Line 34. If some information is not mentioned by the user, it can be mentioned by the system at a later point of the dialogue.
By incorporating system moves that implement the listed rhetorical choices, the modelled dialogue is expected to offer good possibilities for adaptation to high user involvement. Chapter 7 presents a user study that evaluates if the chosen rhetorical style for high user involvement increase the perceived naturalness of the resulting dialogue.

6.4 Functional Testing

The dialogue described in the previous section is utilised to assert the functionality of the implemented dialogue system. For the dialogue system to be considered both multimodal and adaptive, the following criteria have to be met:

- The user move, as well as the level of user involvement have to be detected and correctly fused.
- A system move has to be chosen taking into account the level of user involvement.
- The output has to be provided by the means of an animated agent as well as speech.

The functionality of the implemented multimodal, adaptive dialogue system was tested by two participants uninvolved in the implementation process. One of them was instructed to read a book while talking to the agent in order to simulate an uninvolved user behaviour. To induce user involvement, the second participant was told that they would have to answer questions about the dialogue afterwards. Both users were requested to get information about how to increase the fluid intake of a caretaker.
In both cases, the dialogue system was able to correctly classify the involvement level, fuse it with a user statement and adapt the dialogue strategy in the intended way. The output was conveyed correctly by an animated agent as well as speech. Therefore, it can be asserted that multimodality as well as adaptivity are functioning correctly.

Modelled Dialogue

To conclude this chapter, a graphical representation of the modelled dialogue is provided in Figures 6.2-6.14. It contains system and user moves in different colours, that are labelled with their respective semantic content. Directed edges indicate possible sequences, whereby coloured edges require a high user involvement.

Due to the complexity of the dialogue, its representation extends to several figures connected by special two-coloured nodes. End nodes point to the figure that continues the dialogue flow and indicate by their main colour, which kind of dialogue move follows. Smaller start nodes mark entry points and identify the previous kind of dialogue move.

Furthermore, to avoid too many overlapping edges, dialogue moves may be grouped. Edges to and from grouping nodes apply to all incorporated nodes.

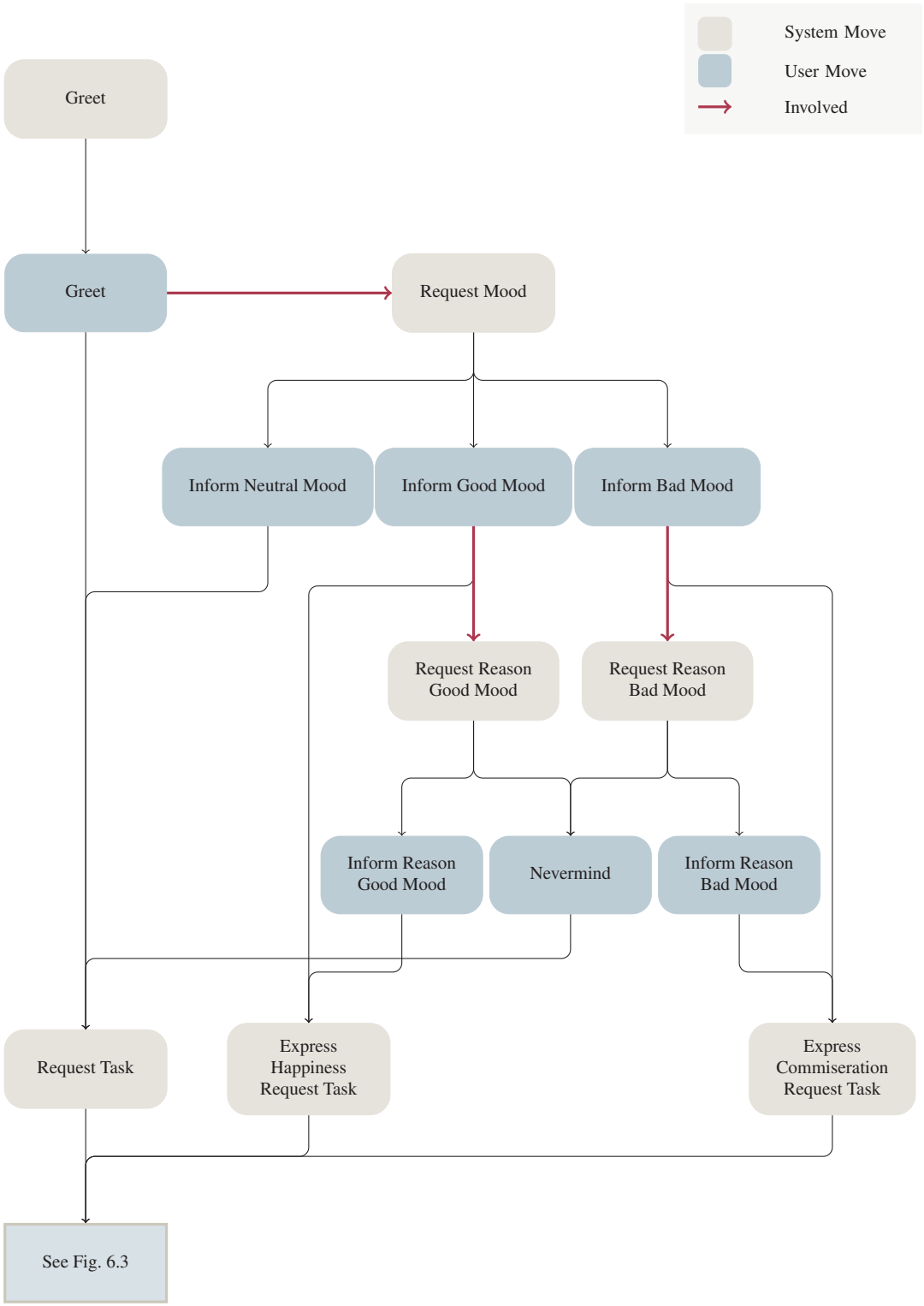

Figure 6.2: A graphical representation of the first part of the dialogue modelled in the scope of this work. It illustrates the initial greetings.

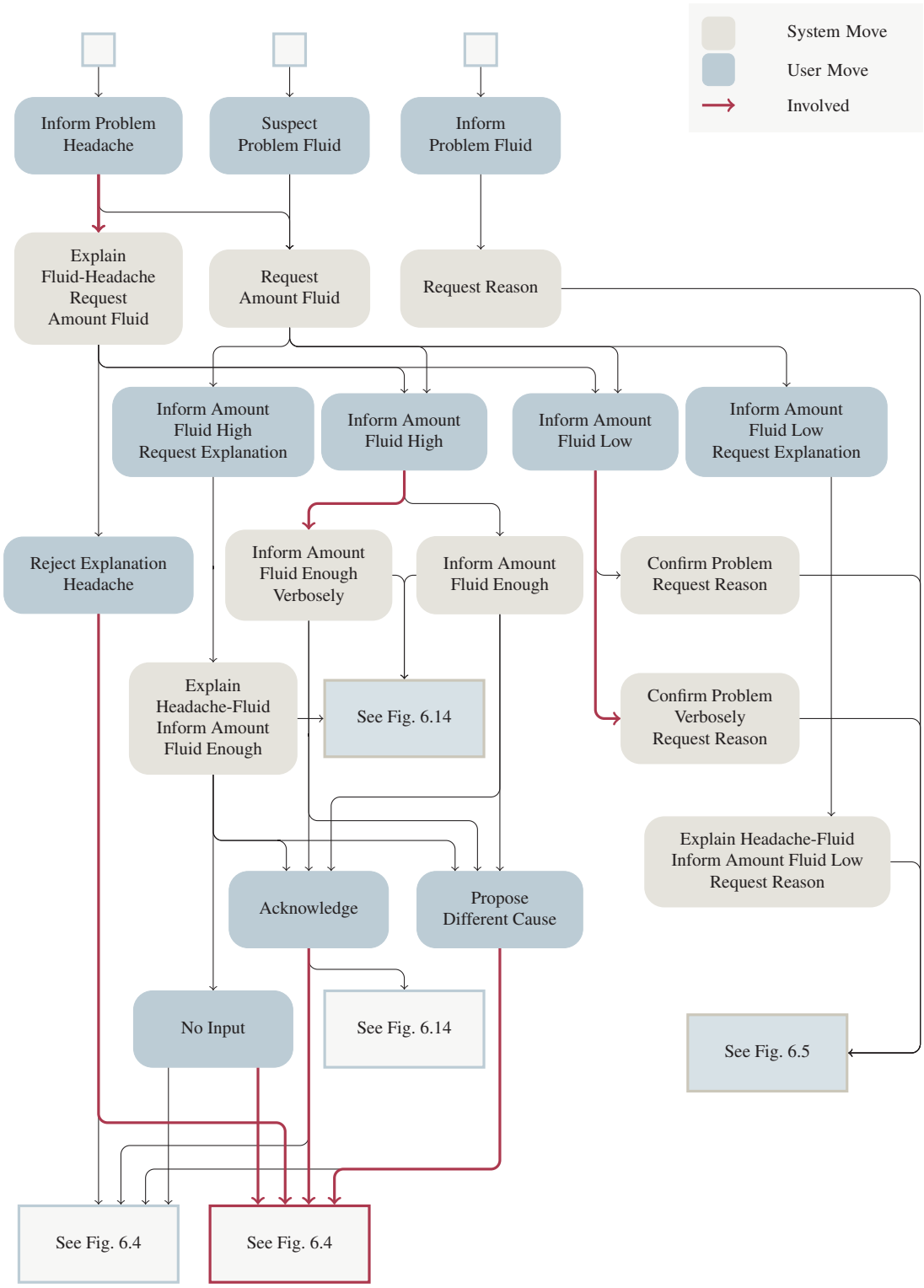

Figure 6.3: A graphical representation of a part of the dialogue modelled in the scope of this work. It deals with determining the problem of the user.

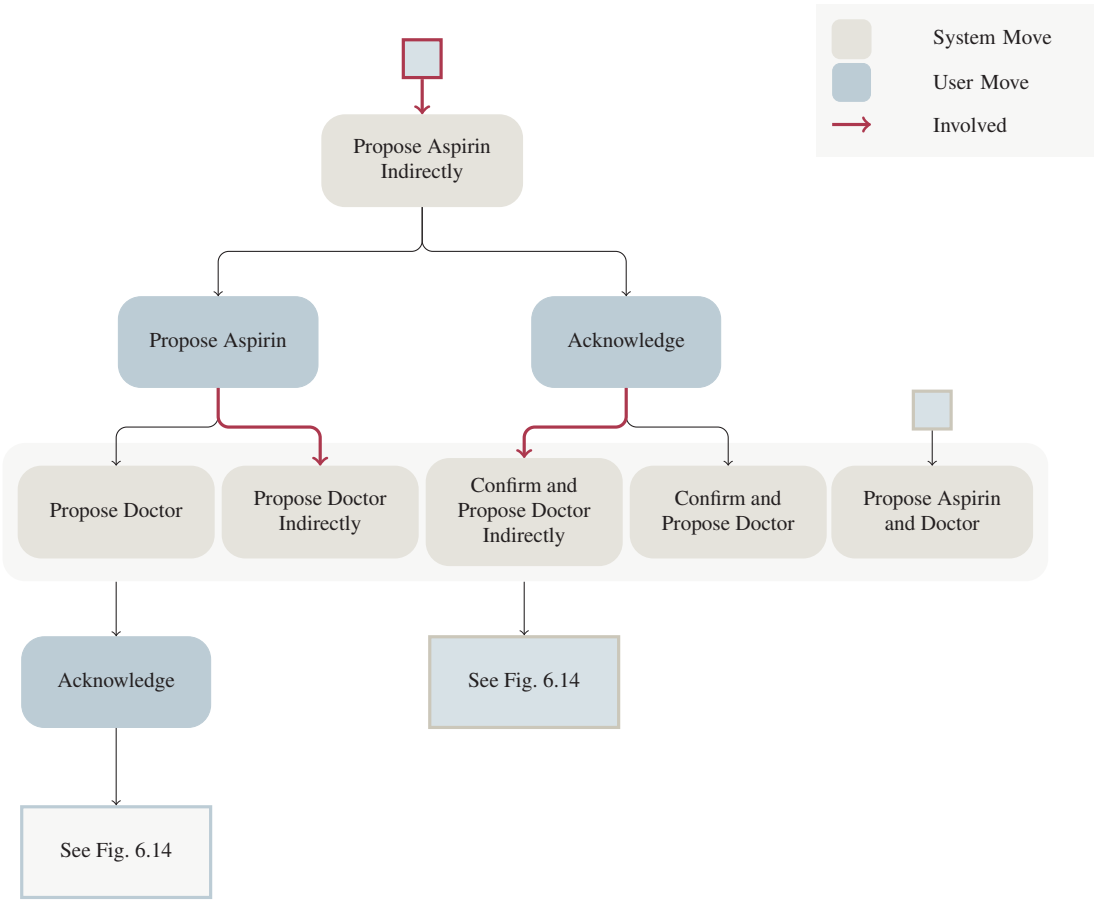

Figure 6.4: A graphical representation of a part of the dialogue modelled in the scope of this work. It deals with the treatment of a headache.

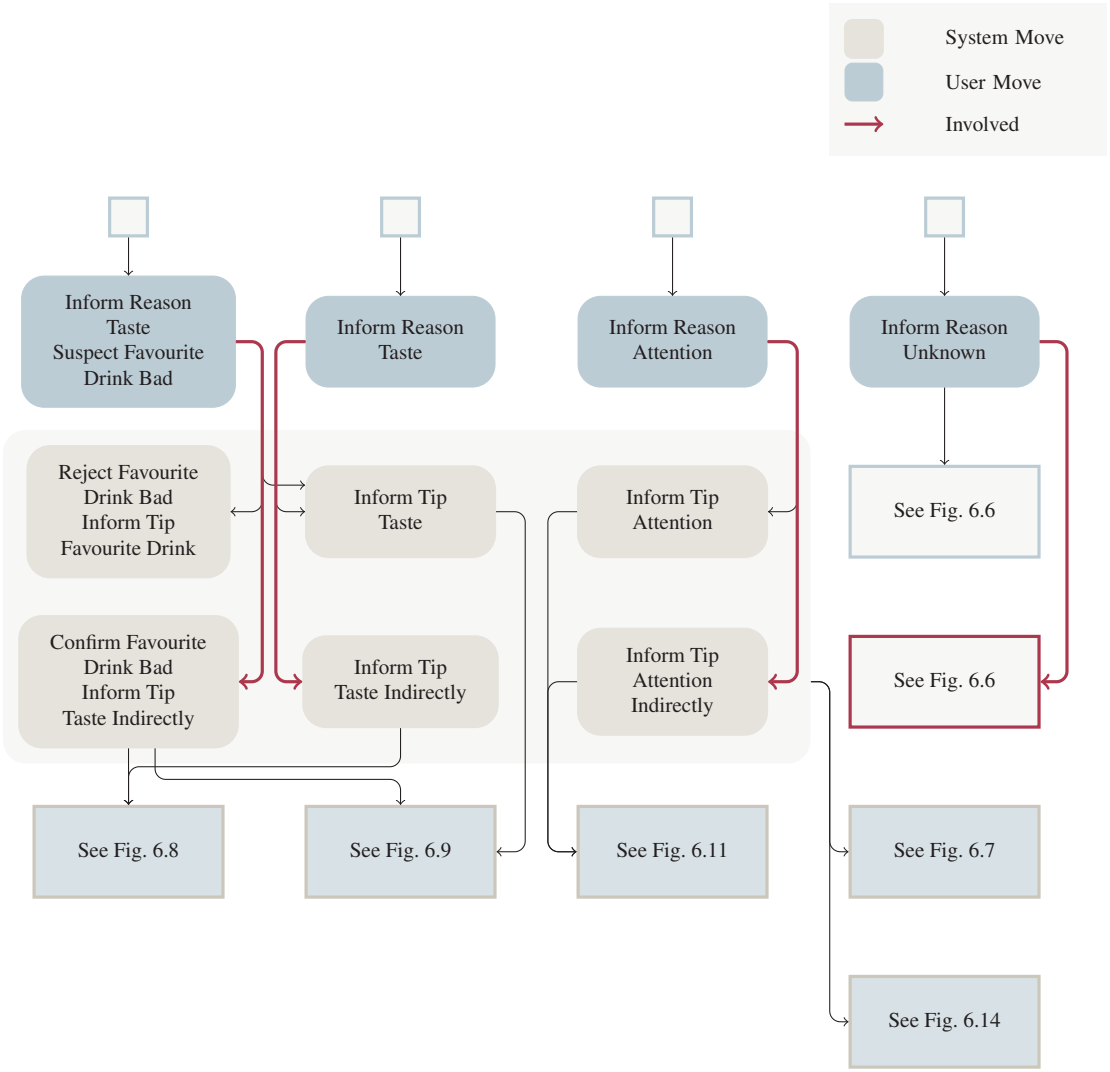

Figure 6.5: A graphical representation of a part of the dialogue modelled in the scope of this work. It deals with determining the reason for a low fluid intake by the patient.

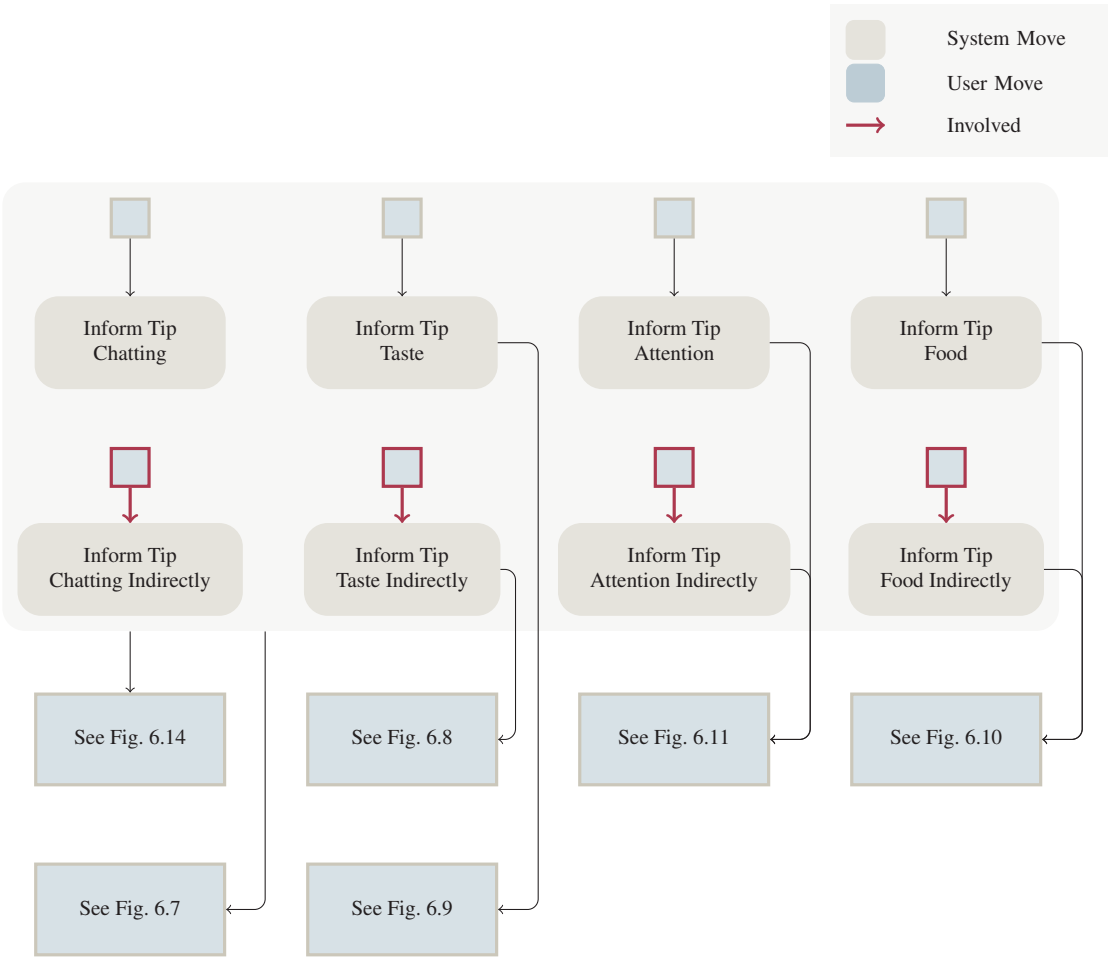

Figure 6.6: A graphical representation of a part of the dialogue modelled in the scope of this work. It list possible ways to increase the fluid intake of a patient.

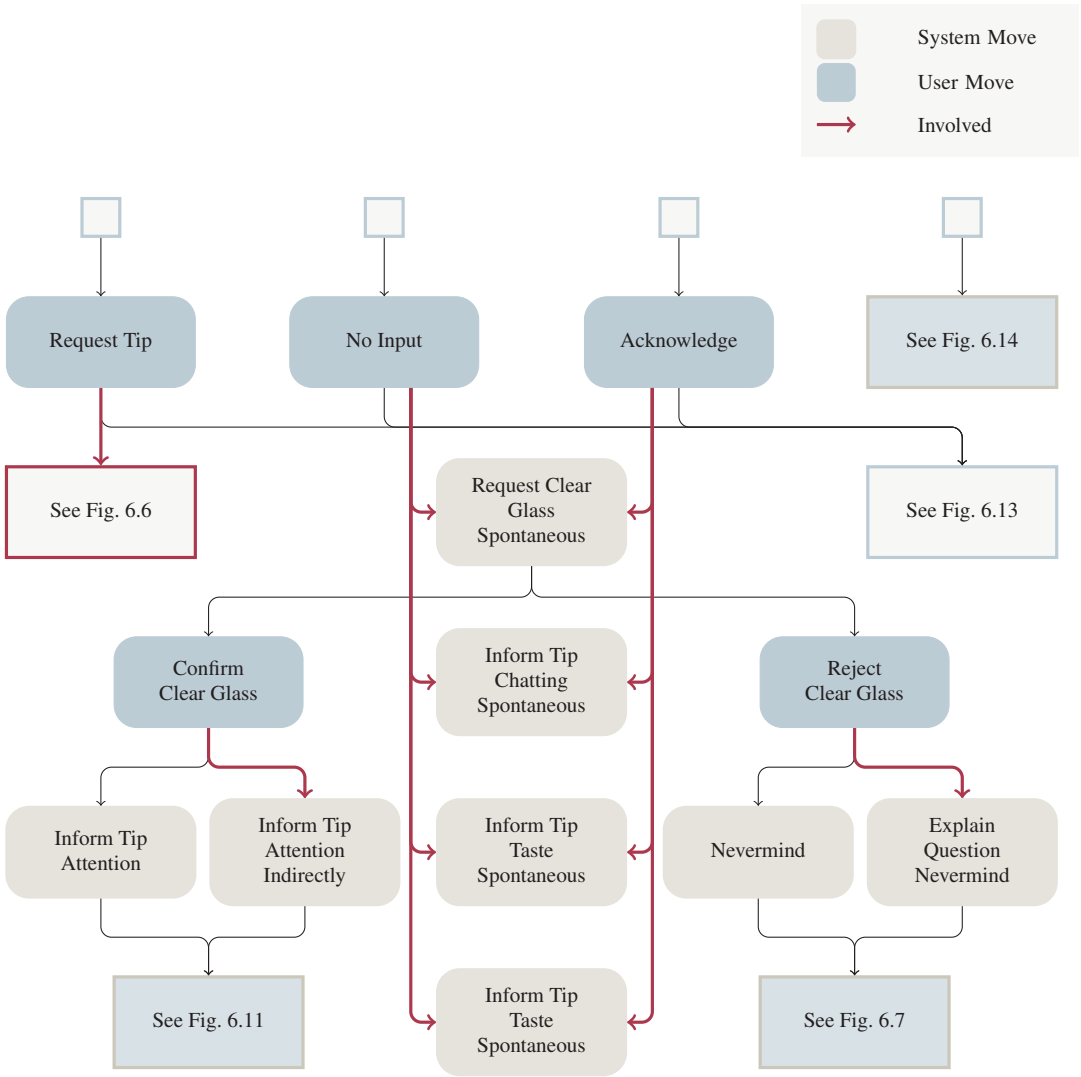

Figure 6.7: A graphical representation of a part of the dialogue modelled in the scope of this work. It contains possible ways to increase the fluid intake of a patient, presented to the user in a spontaneous way.

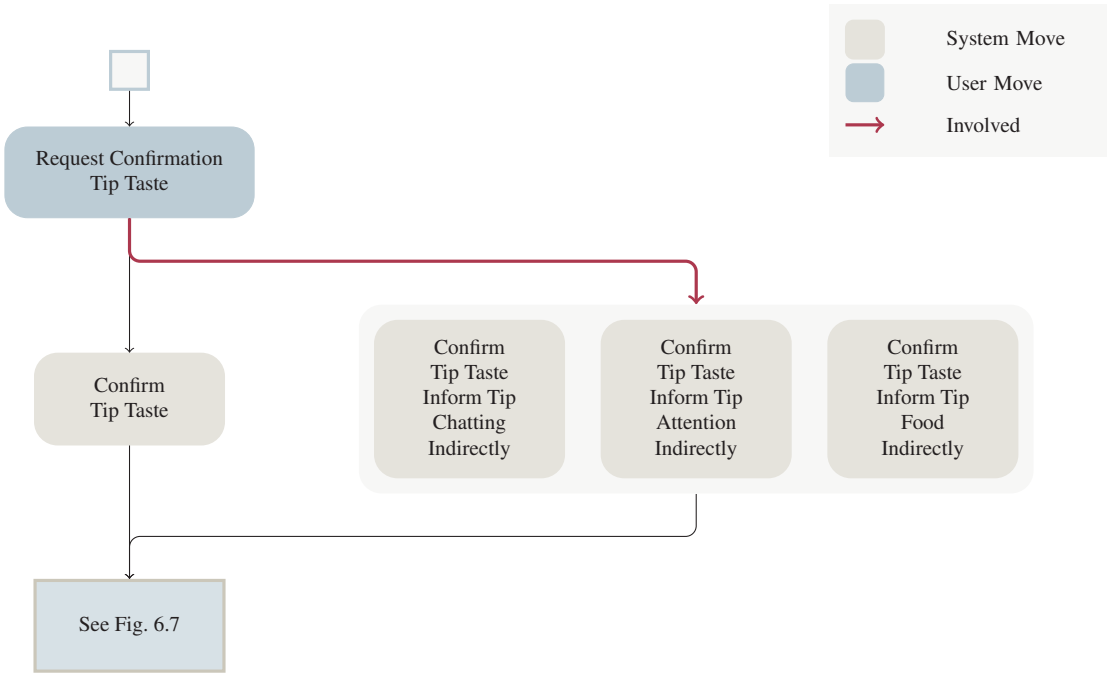

Figure 6.8: A graphical representation of a part of the dialogue modelled in the scope of this work. It deals with confirming the user's impression of the tip given and possibly adding another tip.

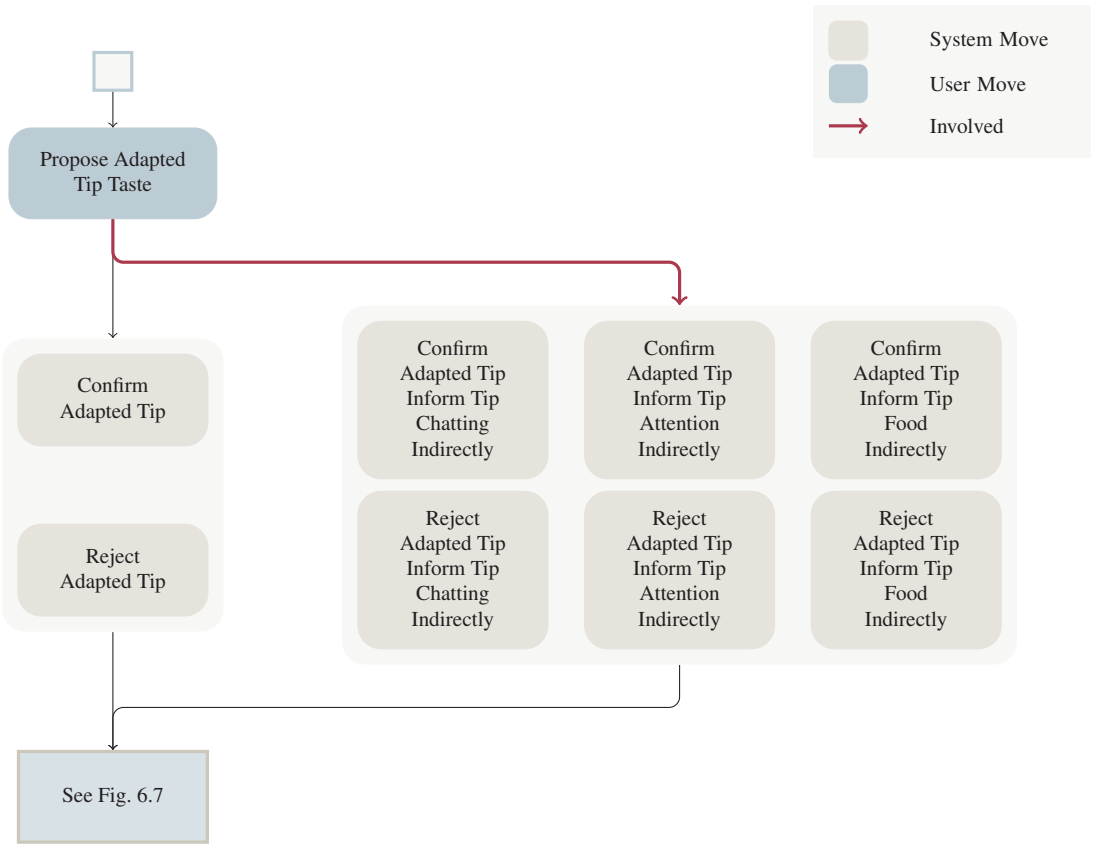

Figure 6.9: A graphical representation of a part of the dialogue modelled in the scope of this work. It handles adaptations by the user of a tip given by the system.

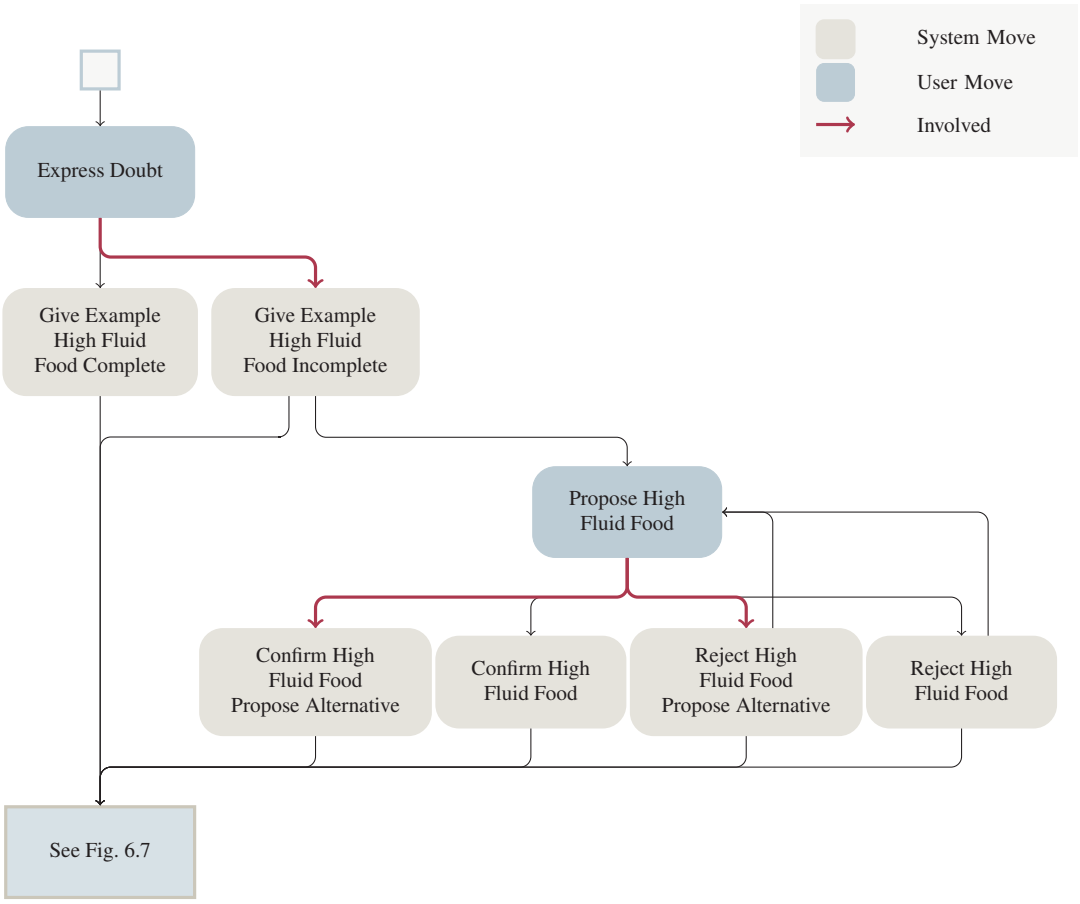

Figure 6.10: A graphical representation of a part of the dialogue modelled in the scope of this work. The system proposes examples of food suitable to increase the fluid intake.

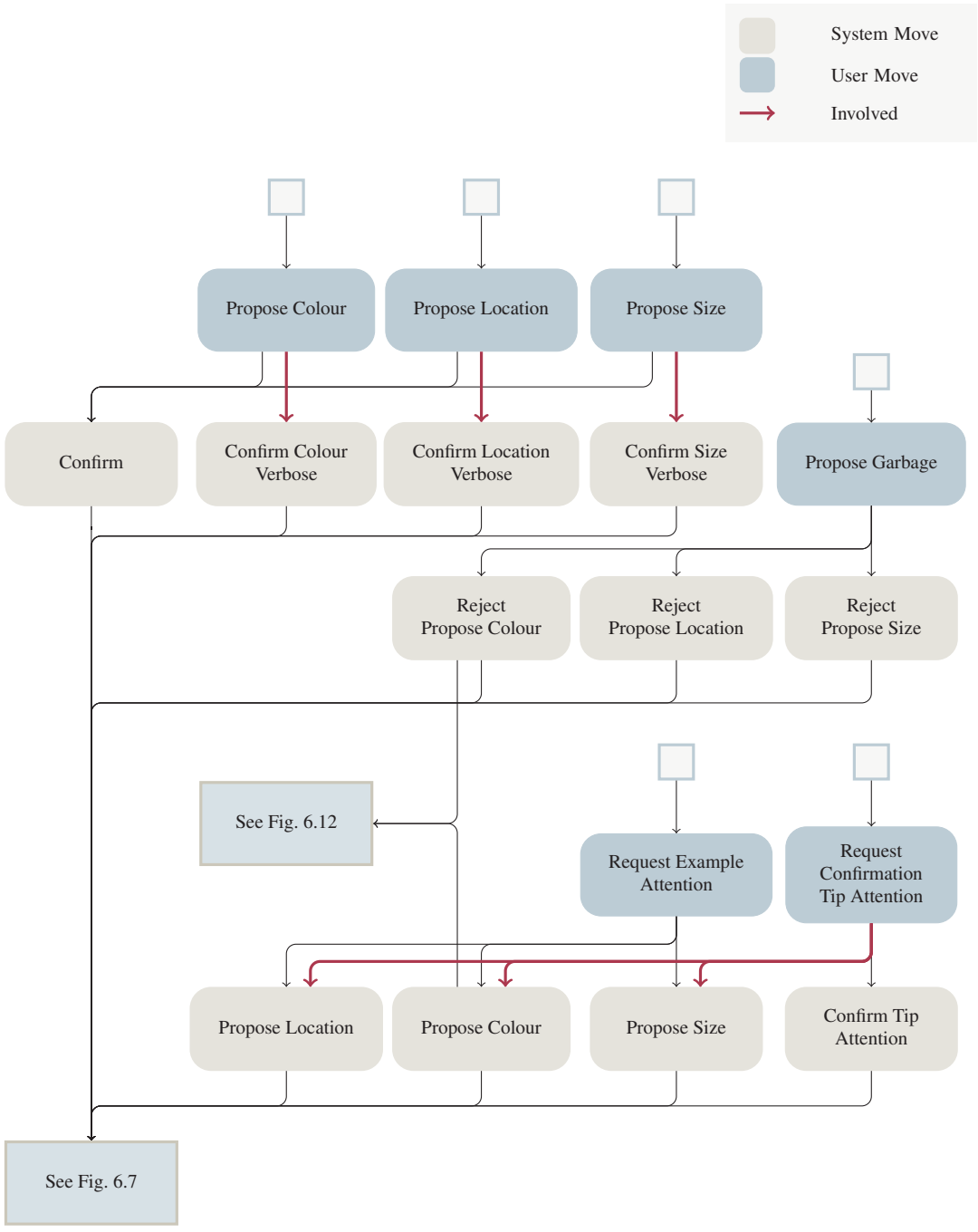

Figure 6.11: A graphical representation of a part of the dialogue modelled in the scope of this work. Different ways are discussed to direct the patient's attention to drinking something.

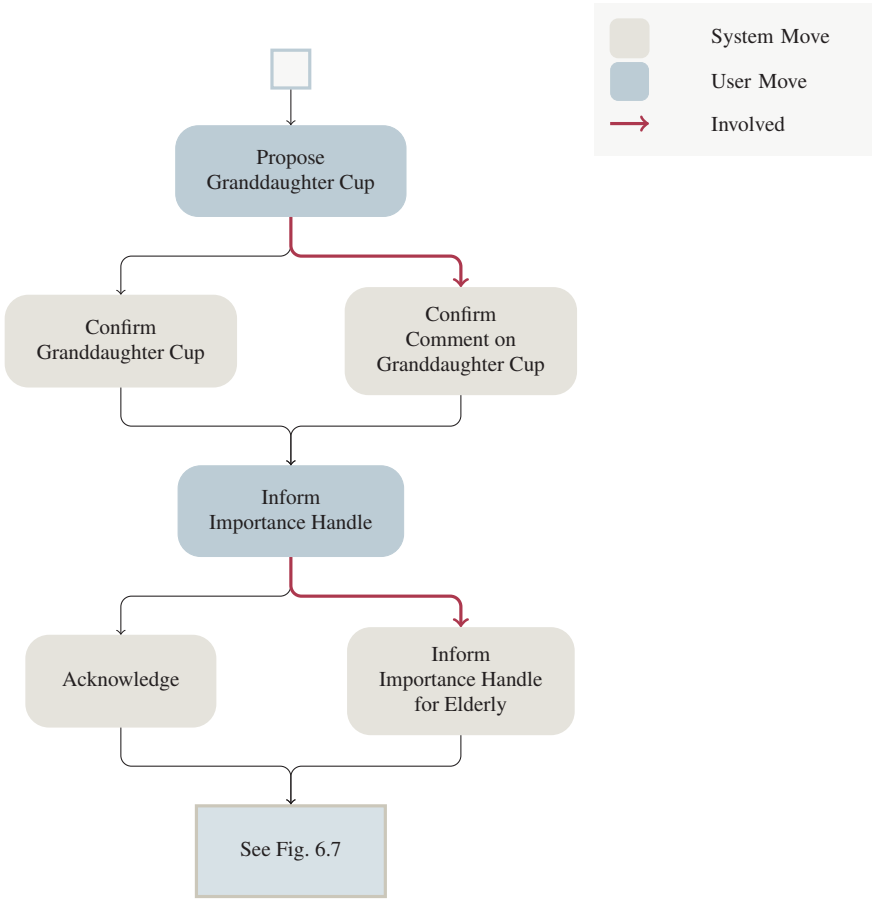

Figure 6.12: A graphical representation of a part of the dialogue modelled in the scope of this work. This part incorporates dialogue steps that do not directly relate to the problem. This kind of conversation is difficult to model in detail for all possible topics that may arise.

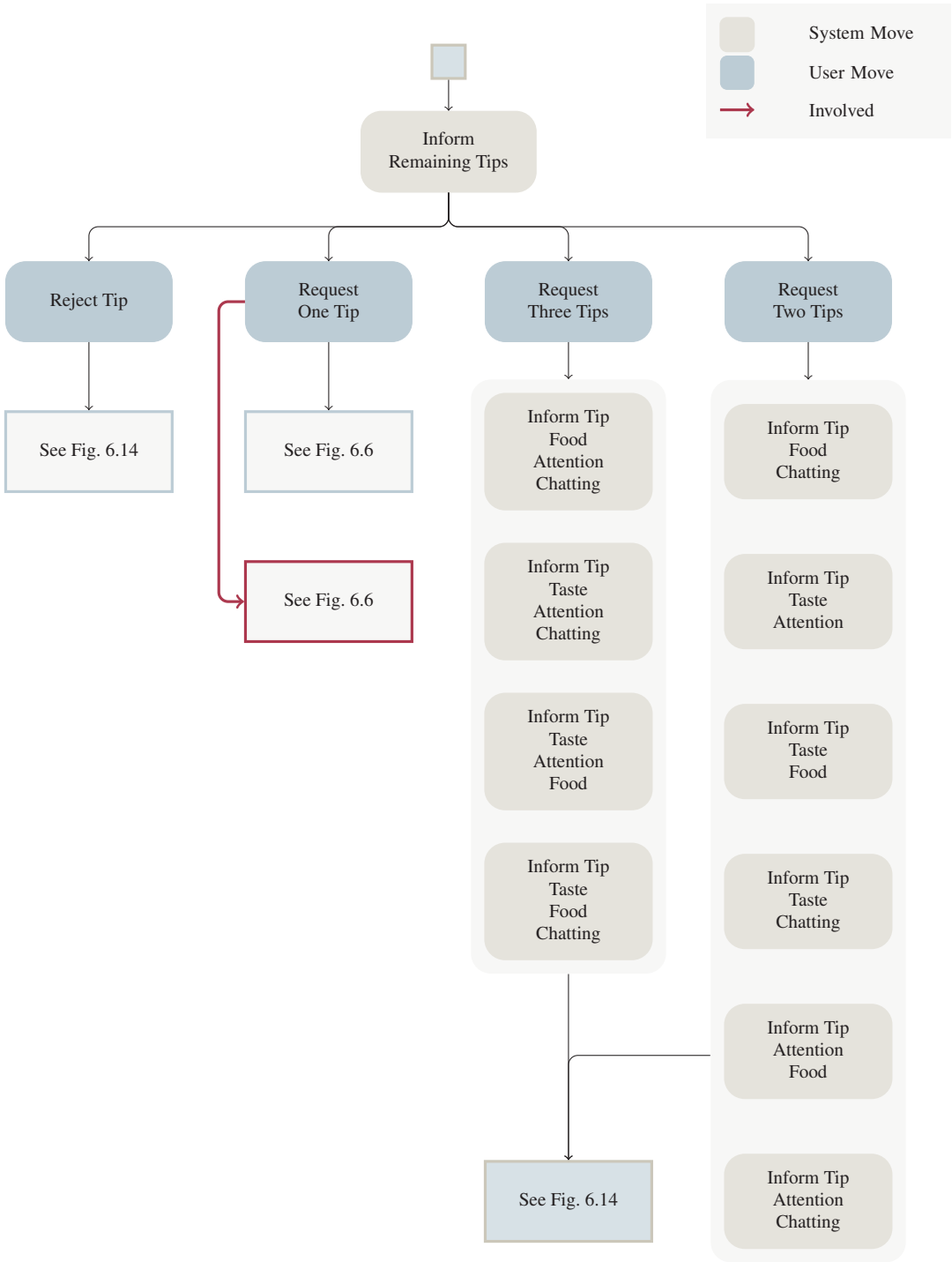

Figure 6.13: A graphical representation of a part of the dialogue modelled in the scope of this work. Tips to increase the patient's fluid intake are provided in a structured way. The user knows what to expect at any time.

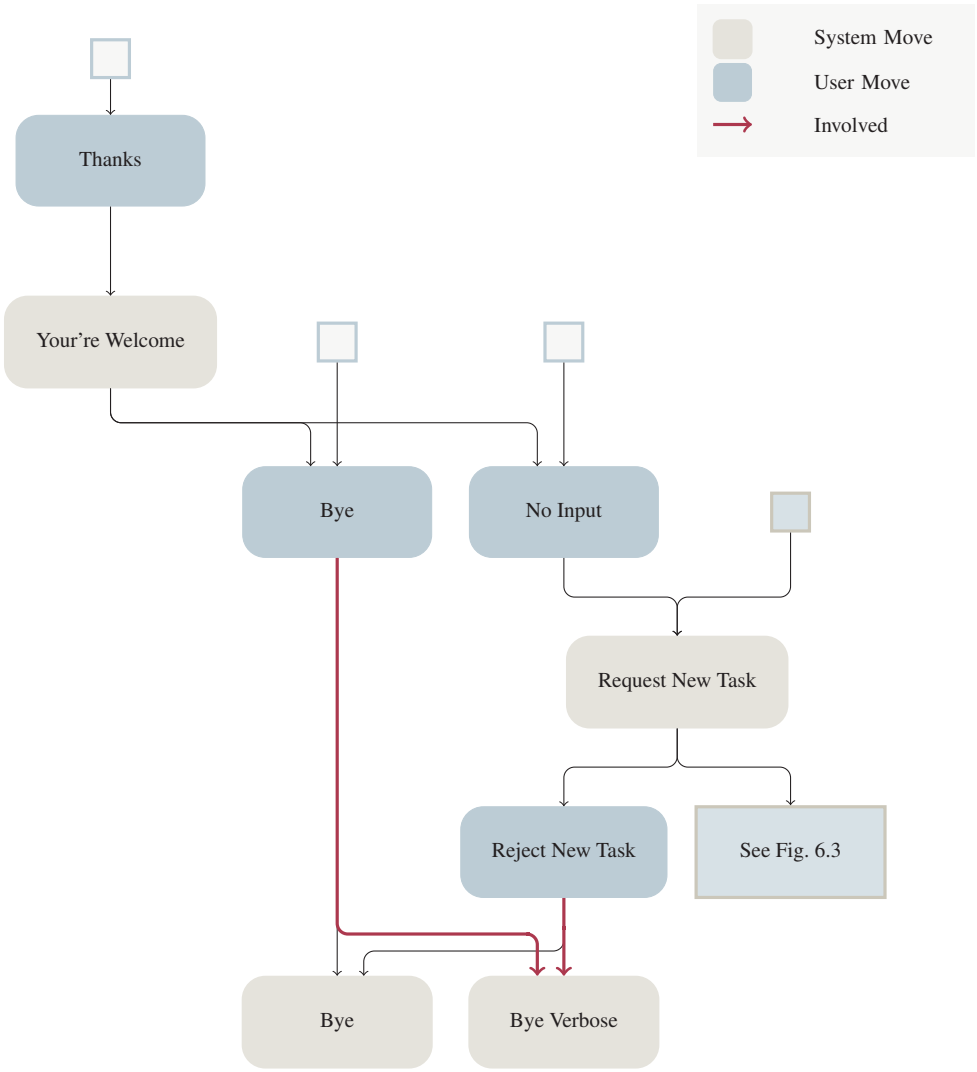

Figure 6.14: A graphical representation of a part of the dialogue modelled in the scope of this work. It handles the leave taking.

7 User Study Regarding the Impact of Rhetorical Style

As discussed in Chapter 4, different rhetorical styles can be observed across cultures in human-human communication, and it is reasonable to assume that emotions can be expressed by rhetorical style. A user study is conducted to evaluate the impact of rhetorical style on the perceived naturalness of a dialogue, as well as on the user preference. The results of this study indicate whether rhetorical style should be considered a relevant factor in future dialogue management systems.
In the following, the setup of the user study is described in detail, before presenting the results and their implications for dialogue management.

7.1 Setup

The goal of the conducted user study was to evaluate the impact of rhetorical style on the perceived naturalness of a dialogue, as well as on the user preference. An online questionnaire was utilised in this endeavour. Participants witnessed two recorded dialogues with different rhetorical styles, and were asked to rate the course of the dialogues on a five-point rating scale.
In contrast to a user study that lets participants interact with the dialogue system, using an online questionnaire, participants do not have first-hand experience with the dialogue system. However, this approach ensures that all participants experience and assess the same interaction. Furthermore, the better availability results in an increased number of participants. Therefore, a decision was made in favour of the online questionnaire.
The depicted scenario in both videos is the similar: a caregiver, called Louisa, interacts with the dialogue system by an virtual avatar called Christian. She suspects that her patient Mr. Smith does not drink enough and expects help from the dialogue system. The dialogues in the videos vary only in regard to the rhetorical style used by the system. The length of the dialogues, the proportion of contributions by Christian and Louisa, the information given by Christian and the input of Louisa are almost identical, in order to ensure that detected differences are due to the rhetorical style.
The independent variable of the user study is the *rhetorical style* of the dialogue, instantiated with the two levels *concise* and *verbose*. The *concise* level embodies the common approach of dialogue systems to rhetorical style : planned and linear reasoning as well as short, precise and direct answers. In contrast, the *verbose* level is characterised by a seemingly spontaneous argumentative structure and answers that can be indirect, incomplete or elaborate. For each level, a dialogue between a user and the dialogue system was recorded. A transcription of the dialogues can be found in Dialogues 6.1 and 6.2. The order in which the videos were presented to the participants was randomized.
Although participants were asked to rate exclusively the course of the dialogue and to ignore factors such

SYSTEM: Hello <User>.
USER: [*involved*] Hello Kristina.
SYSTEM: How can I help you?
USER: [*involved*] I'm afraid Mr. Smith doesn't drink enough during the day.
SYSTEM: How much does Mr. Smith usually drink?
USER: [*involved*] Only one cup of tea after breakfast.
SYSTEM: You're right, that is not enough. Do you know why he doesn't drink enough?
USER: [*involved*] He doesn't like the taste of water. He would rather have some soft drink, but I don't think that is good for him.
SYSTEM: Then you should try sweetened tea. It tastes good and is not as bad as a soft drink.
USER: [*involved*] Perhaps I don't even need to sweeten it. He does drink tea in the morning, after all.
SYSTEM: Exactly. A little company may help as well. While sitting and chatting you can easily drink some glasses without even noticing it.
USER: [*involved*] I guess that's true. Chatting is just more comfortable with a nice cup of tea.
SYSTEM: By the way, does Mr. Smith use a clear glass?
USER: [*involved*] Yes, he does. A beer glass with a handle, so he can better grab it. Why?
SYSTEM: That might be too unobtrusive. A clear glass filled with water does not really attract attention.
USER: [*involved*] Do you think something more eye-catching would be better.
SYSTEM: It is worth a try. Perhaps something more colourful?
USER: [*involved*] I could use the cup his granddaughter gave him. It's very colourful.
SYSTEM: That is a very good idea. He will be happy whenever he sees that cup.
USER: [*involved*] I think so, too. He loves her very much. Also, that cup has a handle. That is very important for him, he feels more secure when grabbing something with a handle.
SYSTEM: Many elderly feel that way.
USER: [*involved*] But what if Mr. Smith still doesn't drink enough? After all, I need to increase that quite a lot.
SYSTEM: The body doesn't only get fluid by drinking. Food contains fluid too.
USER: [*involved*] I don't think that will be enough.
SYSTEM: There some dishes that contain a lot of fluid, such as soup.
USER: [*involved*] Or goulash? He really likes that.
SYSTEM: Yes, goulash is fine. And fruits.
USER: [*involved*] Alright. Thank you, Kristina. You really helped me.
SYSTEM: No problem at all.
USER: [*involved*] I'll get back to work.
SYSTEM: I will be here if you need me.

Dialogue 7.1: The dialogue used in the user study as rhetorical style *verbose*.

SYSTEM: Hello <User>.
USER: [*involved*] Hello Kristina.
SYSTEM: How can I help you?
USER: [*involved*] I'm afraid Mr. Smith doesn't drink enough during the day.
SYSTEM: How much does Mr. Smith usually drink?
USER: [*involved*] Only one cup of tea after breakfast.
SYSTEM: You're right, that is not enough. Do you know why he doesn't drink enough?
USER: [*involved*] He doesn't like the taste of water. He would rather have some soft drink, but I don't think that is good for him.
SYSTEM: Then you should try sweetened tea. It tastes good and is not as bad as a soft drink.
USER: [*involved*] Perhaps I don't even need to sweeten it. He does drink tea in the morning, after all.
SYSTEM: Yes, you can try that.
USER: [*involved*] But what if Mr. Smith still doesn't drink enough? After all, I need to increase that quite a lot.
SYSTEM: I have got three more ideas. Do you want to hear all of them?
USER: [*involved*] Yes, please.
SYSTEM: First, you should keep him company. While chatting people usually drink a lot without even realising it.
USER: [*involved*] I guess that's true. Chatting is just more comfortable with a nice cup of tea.
SYSTEM: Second, you can increase his fluid intake by preparing dishes that contain much fluid, such as soup, goulash and fruits.
USER: [*involved*] He really likes goulash.
SYSTEM: Does Mr. Smith use a clear glass?
USER: [*involved*] Yes, he does. A beer glass with a handle, so he can better grab it. Why?
SYSTEM: In that case you should give him a more colourful glass. A clear glass filled with water does not really attract attention.
USER: [*involved*] I could use the cup his granddaughter gave him. It's very colourful.
SYSTEM: Then it is a good choice.
USER: [*involved*] Alright. Thank you, Kristina. You really helped me.
SYSTEM: No problem at all.
USER: [*involved*] I'll get back to work.
SYSTEM: I will be here if you need me.

Dialogue 7.2: The dialogue used in the user study as rhetorical style *concise*.

as prosody and animation of the avatar, it is questionable if they were able to achieve this. As those factors can influence the perceived naturalness of the dialogue as well as the user preference, a second independent variable was introduced: *dialogue partner* with the two levels *system* and *human*. For the human condition, the part of the system was played by a human actor instead of the implemented dialogue system. This is expected to increase the overall naturalness of the dialogue and facilitate the comparison between the different rhetorical styles. Participants were assigned to one of the two conditions randomly. For the dependent variables, the questionnaire contains ten questions that were answered by the participants using a five-point rating scale:

Qu01: Is Christian helpful?

Qu02: Is Louisa emotionally involved in the dialogue?

Qu03: Does Christian plan his answers?

Qu04: How responsive is Christian to Louisa's contributions?

Qu05: Are Christian's answers spontaneous?

Qu06: Is Christian emotionally involved in the dialogue?

Qu07: How natural is the course of dialogue?

Qu08: How much would you like to participate in such a dialogue?

Qu09: Which dialogue is more natural?

Qu10: In which conversation would you rather participate?

The first two questions make sure that the recordings satisfy two basic requirements: that the dialogue system is considered helpful in all situations and that the actor portraying Louisa conveys involvement in the dialogue in a satisfactory manner. As the adaptation of the rhetorical style aims at involved users, a user that is perceived as uninvolved could inadvertently influence the ratings. Both questions can be rated from 1 - *not at all* to 5 - *very much*. If the average rating is significantly higher than 3, the requirements are considered to be satisfied. This results in the following hypotheses:

$$mean_{Qu01,verbose} > 3$$
$$mean_{Qu01,concise} > 3$$
$$mean_{Qu02,verbose} > 3$$
$$mean_{Qu02,concise} > 3$$

Questions 3-6 test whether the intended dialogue properties, such as spontaneity, are implemented by the chosen rhetorical style. Wrong assumptions about the relationship of rhetorical style and those properties during the design of the dialogue could explain unexpected results concerning the perceived naturalness of the dialogue as well as the user preference. The rating scales of these questions are labeled from 1 - *not at all* to 5 - *very much*. The assumption is that for Qu03 the concise rhetorical style scores higher than the verbose, while for the remaining questions, the rating of the verbose style is expected be higher than the

concise. The corresponding hypotheses are:

$$mean_{Qu03,verbose} < mean_{Qu03,concise}$$
$$mean_{Qu04,verbose} > mean_{Qu04,concise}$$
$$mean_{Qu05,verbose} > mean_{Qu05,concise}$$
$$mean_{Qu06,verbose} > mean_{Qu06,concise}$$

Furthermore, the human dialogue partner is assumed to appear more natural, therefore the following is hypothesised:

$$mean_{Qu03,human} < mean_{Qu03,system}$$
$$mean_{Qu04,human} > mean_{Qu04,system}$$
$$mean_{Qu05,human} > mean_{Qu05,system}$$
$$mean_{Qu06,human} > mean_{Qu06,system}$$

Finally, Questions 7-10 test the perceived naturalness of the dialogue and the user preference. While Questions 7 and 8 rate each dialogue separately from 1 - *not at all* to 5 - *very much*, Questions 9 and 10 compare the videos directly with each other and are labelled from 1 - *verbose* to 5 - *concise*. For Questions 7 and 8, the hypotheses are:

$$mean_{Qu07,verbose} > mean_{Qu07,concise}$$
$$mean_{Qu08,verbose} > mean_{Qu08,concise}$$
$$mean_{Qu07,human} > mean_{Qu07,system}$$
$$mean_{Qu08,human} > mean_{Qu08,system}$$

For Questions 9 and 10, the following is hypothesised:

$$mean_{Qu09,human} < mean_{Qu09,system}$$
$$mean_{Qu10,human} < mean_{Qu10,system}$$
$$mean_{Qu09,human} < 3$$
$$mean_{Qu09,system} < 3$$
$$mean_{Qu10,human} < 3$$
$$mean_{Qu10,system} < 3$$

The described user study was conducted and the hypotheses were tested using the collected data. The results are presented in the following section.

Dialogue Partner	Male	Female	Not Specified
Human	63	78	6
System	61	87	6
Total	124	165	12

Table 7.1: The frequencies that were reported in the user study for gender.

Dialogue Partner	<20	20-29	30-39	>39	Not Specified
Human	14	125	3	0	5
System	13	130	8	1	2
Total	27	255	11	1	7

Table 7.2: The frequencies that were reported in the user study for age.

Dialogue Partner	Never Used	One time use	Regular Use	Not Specified
Human	61	68	15	3
System	60	59	29	6
Total	121	127	44	9

Table 7.3: The frequencies that were reported in the user study for experience with dialogue systems.

7.2 Results

The online questionnaire was accessible for 16 days. During this time, 301 participants completed the questionnaire, four of which answered every question with 'don't know' and are therefore excluded from the further evaluation. Of the remaining participants, 147 filled in the questionnaire for the *human* dialogue partner and 148 for the *system* dialogue partner. Slightly more females than males participated, as can be found in Table 7.1. Most participants were between 20 and 29 years old (see Table 7.2) and had only a limited amount of experience regarding dialogue systems (see Table 7.3). Mean, standard deviation, and median, as well as first and third quartile for all questions are given in Tables 7.4-7.7. Tables A.1-A.4 in the appendix contain the complete frequency distribution.

The data collected from the rating scales is ordinal, therefore they are more precisely measured by the median than by the mean, indicating the use of non-parametric statistical tests. However, the user study measures both within- and between-subject effects, and therefore a two-way mixed ANOVA is required, which has no non-parametric equivalent. Considering this and that it is common practice among researchers to treat rating scale data as interval, the decision was made to use parametric tests for the statistical analysis. Given the size of the groups in this user study, the central limit theorem can be applied and a normal distribution of the data is assumed. Homogeneity of variances is tested using Levene's test and is assumed unless stated otherwise. Independence of the data is ensured by the design of the user study.

The collected data are used to test multiple hypotheses. To keep the family-wise error smaller than the usual 0.05, α is corrected by 16, the maximum number of hypothesis tests that will be performed. This results in $\alpha = 0.0031$ as the limit for significance.

One-sample t-tests, two-way mixed ANOVA and one-way independent ANOVA were used to test the hypotheses listed in the previous section. In the following, the results of the statistical tests will be presented. For each question, a histogram of the rating frequencies is provided in Figures 7.1-7.9.

Question	Video	Dialogue Partner	Mean	σ	Median	Q1	Q3
Qu01	Concise	Human	4.26	0.71	4	4	5
		System	4.28	0.81	4	4	5
		Total	4.27	0.76	4	4	5
	Verbose	Human	4.25	0.71	4	4	5
		System	4.18	0.81	4	4	5
		Total	4.22	0.76	4	4	5
Qu02	Concise	Human	3.41	1.03	4	3	4
		System	3.72	0.83	4	3	4
		Total	3.57	0.94	4	3	4
	Verbose	Human	3.95	0.96	4	4	5
		System	3.99	0.81	4	4	4
		Total	3.97	0.89	4	4	5

Table 7.4: This table shows mean, standard deviation, and median, as well as the first and third quartile for the questions:
Qu01: Is Christian helpful?
Qu02: Is Louisa emotionally involved in the dialogue?

Qu01: Is Christian helpful?

On average, the user rating for this question is significantly higher than 3 for both rhetorical styles, $t(292) = 27.332, p < 0.001$ for *verbose* ($M = 4.22$, $SE = 0.045$) and $t(290) = 28.328, p < 0.001$ for *concise* ($M = 4.27$, $SE = 0.45$). Given these results, it can be concluded that Christian is considered helpful in both dialogues and the hypotheses are accepted.

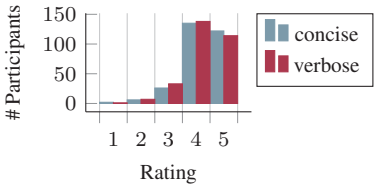

Figure 7.1: Histogram of the rating frequencies for Qu01.

Qu02: Is Louisa emotionally involved in the dialogue?

For this question, both rhetorical styles achieve an average rating that is significantly higher than 3, $t(292) = 18.665, p < 0.001$ for *verbose* ($M = 3.97$, $SE = 0.052$) and $t(289) = 10.207, p < 0.001$ for *concise* ($M = 3.57$, $SE = 0.55$). Given these results, it is likely that the actor of Louisa was able to convey involvement in the dialogue as intended.

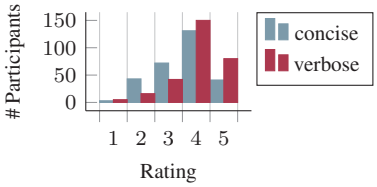

Figure 7.2: Histogram of the rating frequencies for Qu02.

Qu03: Does Christian plan his answers?

There is a significant main effect of the rhetorical style on the rating of the advance planning of Christian, $F(1, 272) = 132.209$, $r = 0.99$, $p < 0.001$. On average, the *concise* style ($M = 4.06$, $SE = 0.06$) received a higher rating than the *verbose* style ($M = 3.18$, $SE = 0.06$). The effect size suggest a large effect. The hypothesis that the concise rhetorical style results in a dialogue that appears to be more planned can therefore be corroborated. In contrast, there was no

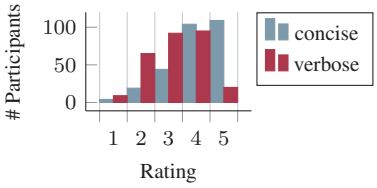

Figure 7.3: Histogram of the rating frequencies for Qu03.

Question	Video	Dialogue Partner	Mean	σ	Median	Q1	Q3
Qu03	Concise	Human	4.03	0.94	4	4	5
		System	4.07	1.01	4	4	5
		Total	4.05	0.97	4	4	5
	Verbose	Human	3.19	1	3	2	4
		System	3.18	0.96	3	2	4
		Total	3.19	0.98	3	2	4
Qu04	Concise	Human	3.55	1.11	4	3	4
		System	3.58	1.03	4	3	4
		Total	3.57	1.07	4	3	4
	Verbose	Human	4.1	0.88	4	4	5
		System	4.29	0.74	4	4	5
		Total	4.19	0.82	4	4	5
Qu05	Concise	Human	2.27	1.04	2	2	3
		System	2.46	0.98	2	2	3
		Total	2.36	1.01	2	2	3
	Verbose	Human	3.22	1.09	3	2	4
		System	3.44	1.02	4	3	4
		Total	3.33	1.06	4	3	4
Qu06	Concise	Human	2.24	1.06	2	1	3
		System	1.85	0.93	2	1	2
		Total	2.04	1.02	2	1	3
	Verbose	Human	2.93	1.15	3	2	4
		System	2.6	1.11	3	2	3
		Total	2.76	1.14	3	2	4

Table 7.5: This table shows mean, standard deviation, and median, as well as the first and third quartile for the questions:
Qu03: Does Christian plan his answers?
Qu04: How responsive is Christian to Louisa's contributions?
Qu05: Are Christian's answers spontaneous?
Qu06: Is Christian emotionally involved in the dialogue?

significant main effect of the dialogue partner on the rating, $F(1, 272) = 0.014$, $p > 0.05$, and also no significant interaction effect between rhetorical style and dialogue partner, $F(1, 272) = 0.038$, $p > 0.05$. On this basis, the second hypothesis regarding this question has to be rejected.

Qu04: How responsive is Christian to Louisa's contributions?

A significant main effect of the rhetorical style on the rating of Christian's responsiveness can be reported, $F(1, 282) = 60.608$, $r = 0.96$, $p < 0.001$. The effect size represents a large effect, with a higher average rating for the verbose rhetorical style ($M = 4.20$, $SE = 0.05$) than for the concise style ($M = 3.57$, $SE = 0.06$). The first hypothesis of this question can therefore be accepted. However, there was no significant main effect of the dialogue partner on the rating, $F(1, 282) = 2.246$, $r = 0.13$, $p > 0.05$, and also no significant interaction effect between rhetorical style and dialogue partner, $F(1, 282) = 1.256$, $p > 0.05$. Although a small effect can be

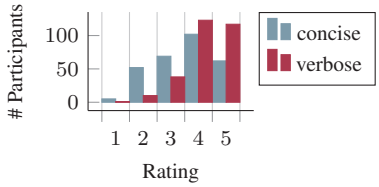

Figure 7.4: Histogram of the rating frequencies for Qu04.

Question	Video	Dialogue Partner	Mean	σ	Median	Q1	Q3
Qu07	Concise	Human	2.66	1.16	2	2	4
		System	2.91	0.95	3	2	4
		Total	2.78	1.07	3	2	4
	Verbose	Human	3.46	1.01	4	3	4
		System	3.7	0.95	4	3	4
		Total	3.58	0.99	4	3	4
Qu08	Concise	Human	2.47	1.13	2	2	3
		System	2.7	1.16	3	2	4
		Total	2.59	1.15	2	2	3
	Verbose	Human	2.99	1.09	3	2	4
		System	3.02	1.11	3	2	4
		Total	3.01	1.1	3	2	4

Table 7.6: This table shows mean, standard deviation, and median, as well as the first and third quartile for the questions:
Qu07: How natural is the course of dialogue?
Qu08: How much would you like to participate in such a dialogue?

reported for the independent variable dialogue partner, the result is not significant and the second hypothesis of this question is rejected.

Qu05: Are Christian's answers spontaneous?

Regarding the perceived spontaneity of Christian, there is a significant main effect of the rhetorical style on the rating, $F(1, 276) = 133.032$, $r = 0.99$, $p < 0.001$. Again, the effect size indicates a large effect and the *verbose* style ($M = 3.33$, $SE = 0.06$) yields higher ratings than the *concise* style ($M = 2.365$, $SE = 0.06$) on average. Hence, the first hypothesis can be accepted. There was no significant main effect of the dialogue partner on the rating, $F(1, 276) = 6.804$, $r = 0.38$, $p > 0.0031$ and also no significant interaction effect between rhetorical style and dialogue partner, $F(1, 276) = 0.145$, $p > 0.05$. The effect size of the main effect of the dialogue partner suggests a medium effect, and with $p = 0.01$ the effect would have been significant if α had not been corrected to reduce the family-wise error. Nevertheless, the second hypothesis regarding this question has to be rejected.

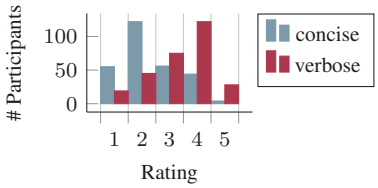

Figure 7.5: Histogram of the rating frequencies for Qu05.

Qu06: Is Christian emotionally involved in the dialogue?

For the emotional involvement of Christian in the dialogue, there is a significant main effect of the rhetorical style, $F(1, 280) = 85.476$, $r = 0.98$, $p < 0.001$, as well as a significant main effect of the dialogue partner on the rating, $F(1, 280) = 12.503$, $r = 0.60$, $p < 0.001$. No significant interaction effect is found between rhetorical style and dialogue partner, $F(1, 280) = 0.102$, $p > 0.05$. Both main effect sizes indicate a large effect. For the independent variable rhetorical style, the *ver-*

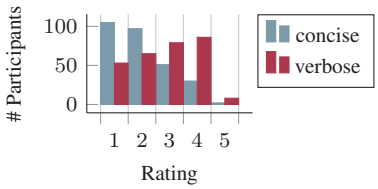

Figure 7.6: Histogram of the rating frequencies for Qu06.

Question	Dialogue Partner	Mean	σ	Median	Q1	Q3
Qu09	Human	2.18	1.27	2	1	3
	System	2.11	1.22	2	1	2
	Total	2.14	1.24	2	1	3
Qu10	Human	2.51	1.29	2	1	3
	System	2.56	1.35	2	1	4
	Total	2.53	1.32	2	1	4

Table 7.7: This table shows mean, standard deviation, and median, as well as the first and third quartile for the questions:
Qu10: Which dialogue is more natural?
Qu09: In which conversation would you rather participate?

bose condition ($M = 2.76$, $SE = 0.07$) achieved a higher average rating than the *concise* condition ($M = 2.04$, $SE = 0.06$). For dialogue partner, *human* ($M = 2.58$, $SE = 0.07$) got higher ratings than *system* ($M = 2.22$, $SE = 0.07$) on average. Therefore, both hypotheses can be accepted.

Qu07: How natural is the course of dialogue?

For this question, Levene's test is significant for the *concise* data set, $F(1, 290) = 11.711$, $p = 0.001$, indicating that the assumption of homogeneity of variance is violated. However, ANOVA is robust against such violations when sample sizes are equal. With 145 valid ratings in the human group and 147 in the system group, the test should be sufficiently reliable. There is a significant main effect of the rhetorical style on the rating of the naturalness of the dialogue, $F(1, 290) = 101.907$, $r = 0.99$, $p < 0.001$. The effect size indicates a large effect and the *verbose* dialogue ($M = 3.58$, $SE = 0.06$) gets higher ratings for naturalness than the *concise* dialogue ($M = 2.78$, $SE = 0.06$) on average. This corroborates the hypothesis that the verbose rhetorical style influences the perceived naturalness of this dialogue positively compared to the concise style. There was no significant main effect of the dialogue partner on the rating, $F(1, 290) = 7.638$, $r = 0.41$, $p > 0.0031$ and also no significant interaction effect between the rhetorical style and dialogue partner, $F(1, 290) = 0.001$, $p > 0.05$. The second hypothesis regarding this question has to be rejected. However, similar to the results of Qu05, the main effect size of the dialogue partner suggests a medium effect, and with $p = 0.006$ the main effect would have been significant, if α had not been corrected to reduce the family-wise error.

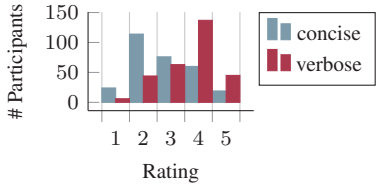

Figure 7.7: Histogram of the rating frequencies for Qu07.

Qu08: How much would you like to participate in such a dialogue?

A significant main effect of the rhetorical style on the rating of the willingness to participate in the dialogue can be reported, with $F(1, 290) = 32.136$, $r = 0.88$, $p < 001$. As for the previous questions, the effect size of the rhetorical style suggesets a large effect and the *verbose* condition ($M = 3.00$, $SE = 0.07$) received a higher average

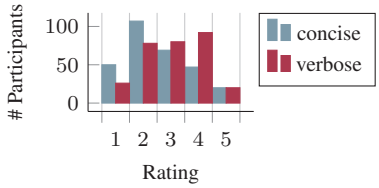

Figure 7.8: Histogram of the rating frequencies for Qu08.

rating than the *concise* condition ($M = 2.59$, $SE = 0.07$). This indicates that the hypothesis can be accepted that the *verbose* rhetorical style increases the participants willingness to participate in the dialogue. Still, there was no significant main effect of the dialogue partner on the rating, $F(1, 290) = 1.450$, $p > 0.05$ and also no significant interaction effect between the rhetorical style and dialogue partner, $F(1, 290) = 2.399$, $p > 0.05$. The second hypothesis has to be rejected.

Qu09: Which dialogue is more natural?

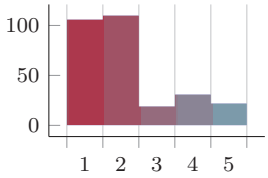

Figure 7.9: Histogram of the rating frequencies for Qu09.

When directly comparing the verbose and the concise dialogue regarding naturalness, no significant difference between the ratings of the group with the system as dialogue partner and the group with the human dialogue partner can be found, $t(286) = 0.454$, $p > 0.05$. Therefore, the hypothesis must be rejected that the rating would be more clearly in favour of the *verbose* rhetorical style if the dialogue partner is human. However, in the collected data, the average rating for this question is significantly different from 3, $t(287) = -11.739$, $r = 0.57$, $p < 0.001$, indicating that the *verbose* dialogue is the more natural one ($M = 2.14$, $SE = 0.07$). Again, the effect size suggests a strong effect of the rhetorical style on the rating. The remaining hypotheses can be accepted.

Qu10: In which conversation would you rather participate?

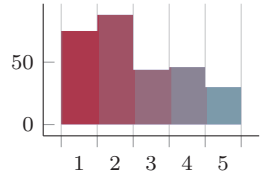

Figure 7.10: Histogram of the rating frequencies for Qu10.

For the direct comparison of the two dialogues regarding the preference for participation, the difference between the rating of the two dialogue partner groups is not significant, $t(281) = -0.332$, $p > 0.05$. Hence, the first hypothesis regarding this question has to be rejected. Regarding the remaining hypotheses, the average rating of this question is significantly different from 3, $t(282) = -5.938$, $r = 0.33$, $p < 0.001$, in favour of the verbose rhetorical style ($M = 2.53$, $SE = 0.8$). The effect size suggests a medium effect of the independent variable on the rating. Therefore, the remaining hypotheses can be accepted.

In conclusion, it can be reported that — contrary to the initial assumption — participants seemed to be able to rate the course of the dialogue without being distracted by disruptive factors such as prosody and animation of the avatar. Almost no differences could be found between the two dialogue partner groups, with the exception of the rating of the emotional involvement of the avatar in the dialogue.

In contrast, the main hypotheses of the user study could be corroborated. The findings strongly suggest that the rhetorical style influences the properties of the dialogue in the way anticipated during the dialogue design, and that in a dialogue such as the one presented, a verbose rhetorical style is perceived as more natural than a concise rhetorical style, and is preferred by the user.

The degree to which the findings can be generalised is limited by the design of the study. An effort was made to exclude potentially interfering factors, for example by using dialogues of approximately the same length, cutting the videos in a way that pauses between conversation contributions have the appropriate length, and asking the participants to rate the course of the dialogue and ignore prosody and animation.

This was done to ensure that observed effects are caused by the variation of rhetorical style. However, it implies that the impact of rhetorical style might be overestimated. In interaction with other factors, the effect of rhetorical style might become smaller in comparison to the other factors that were not considered in this study. Also, the cultural background of the participants was not retrieved. The questionnaire was advertised to university students in Germany. While this indicates that the majority of the participants has a German cultural background and would be expected to prefer the *concise* rhetorical style, it is likely that various participants with other, unknown cultural backgrounds were included. This might influence the results of the conducted user study.

Furthermore, unanswered questions remain. The goal of this user study was to evaluate whether a different rhetorical style than the one most commonly used could be beneficial under specific circumstances. By answering this question, it was intended to determine whether rhetorical style as an adaptation mechanism is a worthwhile research area. Considering the results of the user study, it can be stated that rhetorical style is indeed a promising area of research and should receive further attention. The question under which circumstances which rhetorical style should be used is particularly interesting. The dialogue designed for this user study consisted of a scenario in which using a *verbose* rhetorical style was likely to appear more natural than the more commonly used *concise* one. It cannot be deduced from the findings that the *verbose* rhetorical style is the better choice in all situations. As discussed in Chapter 4, cultural background can have an impact on the preferred rhetorical style. Additionally, further aspects of the emotional state, the situation, or the task may be considered. Further studies have to be conducted in order to determine the interrelation of these factors with the preferred rhetorical style of the user.

The implications of the user study and the aforementioned considerations are that rhetorical style can be considered a promising approach to implement adaptivity. It became apparent that the user's perception of the dialogue is influenced by rhetorical choices of the system, therefore rhetorical style should be explicitly considered during dialogue design. To be able to base design decisions on empirically verified facts, further user studies need to be conducted.

8 Conclusion

This work described extensions of the dialogue manager OwlSpeak regarding adaptivity and multimodality. Computational models of culture and emotion as possible factors of adaptability were discussed to this end. A suitable model for culture stemming from Communication Sciences could be identified. It indicates that changes of the semantic content or the rhetorical style can be utilised for adaptiation to the user's cultural background. Frequently utilised models of emotion stem from psychology and do not offer clear indications for possible adaptations of the system's behaviour. It was hypothesised that variations of rhetorical style can be used as adaptation to emotions, as well.

Following this, the requirements for multimodal dialogue management were deducted by comparing the general architecture of a spoken dialogue system and a multimodal dialogue system. It was concluded that dialogue management is largely unaffected by multimodality, and that changes mostly pertain to the integration of the dialogue manager with the remaining components of a multimodal dialogue system.

The integration of OwlSpeak with Visual SceneMaker to form a fully functional multimodal dialogue system was illustrated, as was the design of a dialogue domain that enables adaptability. The modelled dialogue incorporates variations of rhetorical style as adaptation mechanism for high user involvement.

Finally, a user study was conducted with the help of the developed dialogue system. Two dialogues with different rhetorical styles have been compared using an online questionnaire. The results show that, under the circumstances given in these dialogues, a rhetorical style different from the concise style usually utilised by dialogue systems can be perceived as more natural by the users and be the preferred choice.

The findings of this work lead to many areas of future work. In Chapter 4, computational models for culture and emotion were discussed. It became apparent that it is not yet possible to classify all cultures in regard to the proposed cultural model. Extensive user studies have to be conducted in order to correctly portray the communication patterns of specific cultures in a dialogue system.

Furthermore, communication patterns in reaction to the emotional state of the dialogue partner have received little attention in the literature. Current dialogue systems usually utilize singular system moves in response to a detected emotion. Adaptations of the overall dialogue strategy, for example by changing the rhetorical style, are rarely implemented in existing systems. However, the conducted user study indicates that stylistic choices can positively influence the perceived naturalness and user preference of a dialogue in certain situations. Further user studies should be conducted in order to determine what kind of changes to the dialogue strategy can be utilised in regard to emotions and which dialogue strategy is appropriate for which emotional state.

Regarding emotions, this work focused on the appropriate reaction of the dialogue system to the user's emotional state. However, Butler et al. [6] found that not showing emotions in social encounters can prevent social bonding in human-human interaction. It can therefore be assumed that showing emotions is an important factor for dialogue systems such as the KRISTINA agent, as they aim to create a trust-based

relationship with the user. To enable such behaviour, it is desirable to enhance the OwlSpeak dialogue manager with an emotional state of it own, and to explore dialogue strategies that take into account the system's emotional state.

In Chapter 6, a dialogue for a small domain was modelled, taking into account different rhetorical styles and enabling the dialogue system to respond to some limited off-topic remarks by the user. This results in a complex dialogue model, even for a very limited amount of topics. By connecting OwlSpeak to a knowledge base, the handling of complex dialogue models that are needed for a natural conversation might become more manageable.

In conclusion, this work has demonstrated in what way OwlSpeak can be extended to work in a multimodal, adaptive dialogue system. Suggestions for future work include user studies to further the understanding of communication patterns in human-human communication and the transfer of the gained insights to human-computer communication, as well as the enhancement of OwlSpeak by additional components such as an emotion state of the system and a knowledge base.

A Appendix

Question	Video	Dialogue Partner	1	2	3	4	5
Qu01	Concise	Human	1	2	11	76	55
		System	1	4	15	59	67
		Total	2	6	26	135	122
	Verbose	Human	0	3	14	72	57
		System	1	4	19	66	57
		Total	1	7	33	138	114
Qu02	Concise	Human	3	29	38	53	20
		System	0	14	34	78	21
		Total	3	43	72	131	41
	Verbose	Human	3	11	20	68	44
		System	2	5	22	82	36
		Total	5	16	42	150	80

Table A.1: This table shows the rating frequencies, whereby darker blue indicates a higher frequency, for the questions:
Qu01: Is Christian helpful?
Qu02: Is Louisa emotionally involved in the dialogue?

Question	Video	Dialogue Partner	1	2	3	4	5
Qu07	Concise	Human	20	60	26	28	11
		System	4	54	50	32	8
		Total	24	114	76	60	19
	Verbose	Human	4	26	33	65	18
		System	2	18	30	72	27
		Total	6	44	63	137	45
Qu08	Concise	Human	29	54	33	20	8
		System	21	53	36	27	12
		Total	50	107	69	47	20
	Verbose	Human	15	34	41	49	7
		System	11	44	39	43	13
		Total	26	78	80	92	20

Table A.2: This table shows the rating frequencies, whereby darker blue indicates a higher frequency, for the questions:
Qu07: How natural is the course of dialogue?
Qu08: How much would you like to participate in such a dialogue?

Question	Video	Dialogue Partner	1	2	3	4	5
Qu03	Concise	Human	1	11	22	58	52
		System	3	8	22	46	57
		Total	4	19	44	104	109
	Verbose	Human	5	32	48	45	12
		System	4	33	44	50	8
		Total	9	65	92	95	20
Qu04	Concise	Human	3	29	29	50	32
		System	2	23	40	52	30
		Total	5	52	69	102	62
	Verbose	Human	1	8	19	63	52
		System	0	2	19	60	65
		Total	1	10	38	123	117
Qu05	Concise	Human	35	60	26	19	3
		System	20	62	30	25	1
		Total	55	122	56	44	4
	Verbose	Human	13	24	36	61	10
		System	6	21	39	61	18
		Total	19	45	75	122	28
Qu06	Concise	Human	41	50	29	20	2
		System	64	47	22	10	0
		Total	105	97	51	30	2
	Verbose	Human	22	30	34	54	5
		System	31	35	45	32	3
		Total	53	65	79	86	8

Table A.3: This table shows the rating frequencies, whereby darker blue indicates a higher frequency, for the questions:
Qu03: Does Christian plan his answers?
Qu04: How responsive is Christian to Louisa's contributions?
Qu05: Are Christian's answers spontaneous?
Qu06: Is Christian emotionally involved in the dialogue?

Question	Dialogue Partner	1	2	3	4	5
Qu09	Human	53	50	11	17	11
	System	53	60	8	14	11
	Total	106	110	19	31	22
Qu10	Human	37	43	26	20	14
	System	38	45	18	26	16
	Total	75	88	44	46	30

Table A.4: This table shows the rating frequencies, whereby darker blue indicates a higher frequency, for the questions:
Qu09: Which dialogue is more natural?
Qu10: In which conversation would you rather participate?

Bibliography

[1] E. André, M. Rehm, W. Minker, and D. Bühler. Endowing spoken language dialogue systems with emotional intelligence. In *Affective Dialogue Systems*, pages 178–187. Springer, 2004.

[2] G. Antoniou and F. Van Harmelen. Web ontology language: OWL. In *Handbook on ontologies*, pages 67–92. Springer, 2004.

[3] S. Asteriadis, K. Karpouzis, and S. Kollias. Feature extraction and selection for inferring user engagement in an HCI environment. In *Human-Computer Interaction. New Trends*, pages 22–29. Springer, 2009.

[4] R. Aylett and A. Paiva. Computational modelling of culture and affect. *Emotion Review*, 4(3):253–263, 2012.

[5] R. A. Bolt. *"Put-that-there": voice and gesture at the graphics interface*, volume 14. ACM, 1980.

[6] E. A. Butler, B. Egloff, F. H. Wilhelm, N. C. Smith, E. A. Erickson, and J. J. Gross. The social consequences of expressive suppression. *Emotion*, 3(1):48, 2003.

[7] R. Catizone, A. Setzer, Y. Wilks, et al. Multimodal dialogue management in the COMIC project. In *Proc. Workshop on Dialogue Systems: Interaction, Adaptation and Styles of Management. 10th Conf. of the EACL*, 2003.

[8] P. Ekman and W. V. Friesen. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124, 1971.

[9] C. Elliott, R. J. Adams, and S. Sockalingam. Multicultural toolkit. Toolkit for cross-cultural collaboration. *Awesome library — US office of Minority Affairs, USA*, 1999.

[10] E. Feghali. Arab cultural communication patterns. *International Journal of Intercultural Relations*, 21(3):345–378, 1997.

[11] P. Gebhard and G. Mehlmann. Visual SceneMaker. `https://github.com/SceneMaker/VisualSceneMaker.git`. Accessed on 2015-09-26.

[12] P. Gebhard, G. Mehlmann, and M. Kipp. Visual SceneMaker — a tool for authoring interactive virtual characters. *Journal on Multimodal User Interfaces*, 6(1-2):3–11, 2012.

[13] M. Gnjatović and D. Rösner. Adaptive dialogue management in the NIMITEK prototype system. In *Perception in Multimodal Dialogue Systems*, pages 14–25. Springer, 2008.

[14] C. Goumopoulos, A. Kameas, H. Hagras, V. Callaghan, M. Gardner, W. Minker, M. Weber, Y. Bellik, and A. Meliones. ATRACO: adaptive and trusted ambient ecologies. In *Self-Adaptive and Self-*

Organizing Systems Workshops, 2008. SASOW 2008. Second IEEE International Conference on, pages 96–101. IEEE, 2008.

[15] T. Heinroth and D. Denich. Spoken interaction within the computed world: evaluation of a multitasking adaptive spoken dialogue system. In *Computer Software and Applications Conference (COMPSAC), 2011 IEEE 35th Annual*, pages 134–143. IEEE, 2011.

[16] T. Heinroth, D. Denich, and A. Schmitt. OwlSpeak — adaptive spoken dialogue within intelligent environments. In *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2010 8th IEEE International Conference on*, pages 666–671. IEEE, 2010.

[17] G. H. Hofstede and G. Hofstede. *Culture's consequences: comparing values, behaviors, institutions and organizations across nations*. Sage, 2001.

[18] N. Jaksic, P. Branco, P. Stephenson, and L. M. Encarnação. The effectiveness of social agents in reducing user frustration. In *CHI'06 extended abstracts on Human factors in computing systems*, pages 917–922. ACM, 2006.

[19] D. Jan, D. Herrera, B. Martinovski, D. Novick, and D. Traum. A computational model of culture-specific conversational behavior. In *Intelligent virtual agents*, pages 45–56. Springer, 2007.

[20] R. B. Kaplan. Cultural thought patterns in inter-cultural education. *Language learning*, 16(1-2):1–20, 1966.

[21] S. Larsson and D. R. Traum. Information state and dialogue management in the TRINDI dialogue move engine toolkit. *Natural language engineering*, 6(3&4):323–340, 2000.

[22] S. Mascarenhas, R. Prada, A. Paiva, and G. J. Hofstede. Social importance dynamics: a model for culturally-adaptive agents. In *Intelligent virtual agents*, pages 325–338. Springer, 2013.

[23] S. McGlashan. Towards multimodal dialogue management. In *Proceedings of Twente Workshop on Language Technology*, volume 11. Citeseer, 1996.

[24] A. Mehrabian. Pleasure-arousal-dominance: a general framework for describing and measuring individual differences in temperament. *Current Psychology*, 14(4):261–292, 1996.

[25] Microsoft. Kinect. `https://dev.windows.com/en-us/kinect`. Accessed on 2015-09-26.

[26] Microsoft. Microsoft Speech Platform. `https://msdn.microsoft.com/en-us/library/office/hh361572%28v=office.14%29.aspx`. Accessed on 2015-09-26.

[27] W. Minker. Spoken language dialogue systems overview: Lecture 1. *ENGC 7011: Dialogue Systems, Slides (Ulm, Vic: Ulm University, WT 2013/2014), unpublished lecture*.

[28] M. Oshry, R. Auburn, P. Baggia, M. Bodell, D. Burke, D. C. Burnett, E. Candell, J. Carter, S. McGlashan, A. Lee, et al. Voice extensible markup language (voicexml) 2.1. *W3C Recommendation*, 2007.

[29] T. Partala and V. Surakka. The effects of affective interventions in human–computer interaction. *Interacting with computers*, 16(2):295–309, 2004.

[30] J. Pittermann and A. Pittermann. A data-oriented approach to integrate emotions in adaptive dialogue management. In *Proceedings of the 12th international conference on Intelligent user interfaces*, pages 270–273. ACM, 2007.

[31] R. Plutchik. *Emotion: a psychoevolutionary synthesis*. Harpercollins College Division, 1980.

[32] M. Potel. MVP: Model-View-Presenter the taligent programming model for C++ and Java. *Taligent Inc*, 1996.

[33] J. Sanghvi, G. Castellano, I. Leite, A. Pereira, P. W. McOwan, and A. Paiva. Automatic analysis of affective postures and body motion to detect engagement with a game companion. In *Human-Robot Interaction (HRI), 2011 6th ACM/IEEE International Conference on*, pages 305–311. IEEE, 2011.

[34] A. Schmitt, B. Schatz, and W. Minker. Modeling and predicting quality in spoken human-computer interaction. In *Proceedings of the SIGDIAL 2011 Conference*, pages 173–184. Association for Computational Linguistics, 2011.

[35] M. Schulz. Horde3D. `http://www.horde3d.org/`. Accessed on 2015-09-26.

[36] D. Szafir and B. Mutlu. Pay attention!: designing adaptive agents that monitor and improve user engagement. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 11–20. ACM, 2012.

[37] S. Ultes, H. Dikme, and W. Minker. Dialogue management for user-centered adaptive dialogue. In *Proceedings of the 5th International Workshop On Spoken Dialogue Systems (IWSDS). Springer, January*, 2014.

[38] S. Ultes and W. Minker. HIS-OwlSpeak: a model-driven dialogue manager with multiple control modes. In *Intelligent Environments (IE), 2013 9th International Conference on*, pages 108–115. IEEE, 2013.

[39] J. Wagner, F. Lingenfelser, T. Baur, I. Damian, F. Kistler, and E. André. The social signal interpretation (SSI) framework: multimodal signal processing and recognition in real-time. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 831–834. ACM, 2013.

[40] S. Young, J. Schatzmann, K. Weilhammer, and H. Ye. The hidden information state approach to dialog management. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 4, pages IV–149. IEEE, 2007.

[41] C. Yu, P. Aoki, and A. Woodruff. Detecting user engagement in everyday conversations. In *Eighth International Conference on Spoken Language Processing*, 2004.

[42] R. S. Zaharna. Understanding cultural preferences of arab communication patterns. *Public Relations Review*, 21(3):241–255, 1995.