Adaptive Dialogue Management in Human-Machine Interaction

Dissertation

zur Erlangung des akademischen Grades

Doktoringenieur
(Dr.-Ing.)

angenommen durch die Fakultät für Informatik
der Otto-von-Guericke-Universität Magdeburg

von: M.Sc. Milan Gnjatović
geb. am 10.09.1978 in Belgrad (Serbien)

Gutachter: Prof. Dr. Dietmar Rösner
Prof. Dr. Andreas Wendemuth
Prof. Dr. Elisabeth André

Promotionskolloquium: Magdeburg, den 07.09.2009

Abstract

Research in the domain of affective computing is usually concentrated on detection of emotional user behavior. However, less attention is devoted to the question how to enable dialogue systems to overcome problems in the interaction related to emotional user behavior. We address the latter research question. This thesis makes contributions to adaptive dialogue management in human-machine interaction in the areas of theory, experimental practice, and system development. In this work, we discuss important design considerations and implementation issues in development of an adaptive dialogue management module, and exemplify them for the NIMITEK (Neurobiologically Inspired, Multimodal Intention Recognition for Technical Communication Systems) prototype spoken dialogue system for supporting users while they solve problems in a graphics system. The introduced approach represents an integration of several lines of research: producing and evaluating corpora of affected behavior in human-machine interaction, modeling attentional information on the level of interpreting the user's command, and designing adaptive dialogue strategies.

This research is essentially supported by the NIMITEK corpus of affected behavior in human-machine interaction collected within the reported research. It contains 15 hours of audio and video recordings produced during a Wizard-of-Oz experiment specially designed to induce emotional reactions. Ten native German speakers participated in the experiment. The evaluation and annotation of the NIMITEK corpus with respect to its emotional content demonstrated a satisfying level of ecological validity: the corpus contains recordings of genuine, not acted, emotions that were overtly signaled; it is not oriented to extreme representations of a few emotions only but comprises also expressions of less intense, everyday emotions; emotional expressions of diverse emotions are extended in modality (voice and facial expression) and time. In addition to audio and video recordings of the experimental sessions, all dialogues are transcribed and dialogue acts are annotated.

The dialogue management module in the NIMITEK prototype system dynamically adapts its dialogue strategy according to the current state of the interaction. We model the state of the interaction as a composite of five interaction features: the state of the task, the user's command, the focus of attention, the state of the user, and the history of interaction. Dynamical adaptation includes three distinct but interrelated decision making processes: When to provide support to the user? What kind of support to provide? How to provide support? These reflect three underlying requirements for a dialogue strategy aimed to support the user. First, support should be timely provided, without relying on the assumption that the user will clearly state a need for support. Second, problems in the interaction may be various (e.g., they may relate to the given task, to the interface language, to the emotional state of the user, etc.) and the user should be provided with useful, sufficient and appropriately emphasized information tailored to a particular problem. And third, the manner of providing support should be tailored to meet the user's needs, i.e., it should be in accordance with the emotional state of the user.

Particular emphasis is devoted to the level of naturalness of interaction. We introduce a model of attentional state on the level of the user's command that facilitates processing of more flexibly formulated commands. The model is demonstrated to work well for different syntactic forms of commands (e.g., elliptical commands, verbose commands, context dependent commands, etc.). In addition, we discuss how the dialogue management module handles miscommunication on different levels: the conversational level, the intentional level, and the signal level.

Although we report an implementation of the dialogue management module for a task-specific scenario (i.e., support in solving the Tower of Hanoi puzzle), the introduced concepts are designed to be task-independent.

Zusammenfassung

Die technisch-orientierte Forschung zum Emotion Computing konzentriert sich üblicherweise auf die Erkennung von emotionalem Verhalten eines Benutzers. Die Erhöhung der assistiven Fähigkeiten von Mensch-Maschine-Schnittstellen in technischen Systemen durch angepasste Dialogstrategien sowie die adäquate Behandlung emotionaler Aspekte in der Mensch-Maschine-Interaktion wurden bislang noch nicht umfassend genug erforscht. In dieser Arbeit wurde diese Forschungsfrage betrachtet. Diese Dissertation trägt zum adaptiven Dialogmanagement in der Mensch-Maschine-Interaktion in den Bereichen Theorie, experimentelle und empirische Praxis und System-Entwicklung bei. Es wird über wichtige Aspekte von Design und Entwicklung des Dialogmanagers im NIMITEK-Prototypsystem (Neurobiologisch inspirierte, multimodale Intentionserkennung für technische Kommunikationssysteme) berichtet. Dieser Dialogmanager unterstützt den Benutzer, um verschiedene Probleme in der Kommunikation bei gesprochener Mensch-Maschine-Interaktion zu überwinden. Der vorgestellte Ansatz verwirklicht eine Integration von mehreren Forschungslinien: der Aufbau und die Auswertung von Korpora emotionalen Verhaltens in der Mensch-Machine-Interaktion, die Modellierung des Fokus der Aufmerksamkeit auf der Ebene von Benutzerkommandos und die Entwicklung von adaptiven Dialogstrategien.

Diese Forschung nutzt das NIMITEK-Korpus über emotionales Verhalten in der Mensch-Machine-Interaktion. Dieses im Rahmen der Arbeiten gewonnene Korpus umfasst 15 Stunden Audio- und Video-Aufzeichnungen des Wizard-of-Oz-Experiments (WOZ), welches entworfen wurde, um Emotionen zu induzieren. Zehn Muttersprachler des Deutschen nahmen als Probanden am Experiment teil. Die Auswertung und die Annotation des NIMITEK-Korpus im Hinblick auf emotionalen Inhalt demonstrierten, dass das Korpus als ökologisch valide gelten kann: das Korpus umfasst Aufzeichnungen von echten—nicht von Schauspielern präsentierten—Emotionen, die offen gezeigt werden; es orientiert sich nicht nur an voll entfalteten Emotionen, sondern auch an wenig intensiven emotionalen Ausdrücken, die dafür repräsentativ sind, was im Alltag passiert; die diversen Emotionen wurden mittels verschiedenen Modalitäten (Sprache und Mimik) und über längere Zeiträume ausgedrückt. Zusätzlich zu den Audio- und Videoaufzeichnungen der experimentellen Sitzungen wurden alle Dialoge des WOZ-Experiments transkribiert und Dialogakte annotiert.

Der Dialogmanager im NIMITEK-Prototypsystem passt seine Dialogstrategie entsprechend der Interaktionsituation dynamisch an. Die Interaktionsituation wird als Komposition der fünf folgenden Interaktionsmerkmalen modelliert: der Zustand der Aufgabe, das aktuelle Nutzerkommando, der Fokus der Aufmerksamkeit, der Zustand des Benutzers, und die Historie der Interaktion. Der Dialogmanager nimmt diese Interaktionsmerkmale, um dynamisch zu entscheiden: Wann soll Unterstützung gegeben werden? Welche Art von Unterstützung sollte gegeben werden? Wie sollte Unterstützung gegeben werden? Diese ausgeprägten, aber zusammenhängenden Entscheidungsprozesse reflektieren drei grundlegende Anforderungen für eine Dialogstrategie zur Unterstützung von Benutzern. Die erste Anforderung ist, dass Unterstützung rechtzeitig gegeben werden sollte, ohne sich darauf zu verlassen, dass der Benutzer den Unterstützungsbedarf deutlich signalisieren wird. In der Mensch-Maschine-Interaktion können verschiedene Probleme auftreten (z.B. Probleme bezogen auf die Aufgabe selbst, Verständigungsprobleme, Probleme bezogen auf den emotionalen Zustand des Benutzers, usw.). Die zweite Anforderung ist nun, dass dem Benutzer nützliche, ausreichende, und angemessen hervorgehobene Information gegeben werden sollte. Die dritte Anforderung ist, dass der Stil, in dem Unterstützung gegeben wird, entsprechend dem emotionalen Zustand vom Benutzer angepasst werden sollte.

Ein weiterer Schwerpunkt wurde auf die Natürlichkeit der Mensch-Maschine-Interaktion gelegt. Es wurde ein Modell der Fokusstruktur auf der Ebene von Benutzerkommandos entworfen, welches die Verarbeitung von flexibel formulierten Kommandos ermöglicht. Es wurde gezeigt, dass das Modell für verschiedene syntaktische Formen der Kommandos (z.B. elliptische Äußerungen, verbose Äußerungen, kontextabhängige Äußerungen, usw.) funktioniert. Außerdem wird darauf eingegangen, wie der Dialogmanager Probleme der Miskommunikation auf den verschiedenen Ebenen (die Konversation-Ebene, die Intention-Ebene und die Signal-Ebene) behandelt.

Ausgehend von den entworfenen, grundlegenden Design-Konzepten für eine aufgabenunabhängige Implementierung, wurde in der Arbeit exemplarisch eine aufgabenspezifische Implementierung des Dialogmanagers (anhand der Aufgabe "Türme von Hanoi") vorgestellt.

Acknowledgements

I would like to thank my advisor, Prof. Dr. Dietmar Rösner, for all his support, guidance, kindness and trust. For all of these, I am deeply indebted to him. My thanks go also to my committee members, Prof. Dr. Andreas Wendemuth and Prof. Dr. Elisabeth André, for their friendly attitude, useful advices and strong support.

I wish to thank many contributors and collaborators in my work, particularly Dr.-Ing. Manuela Kunze, Mirko Hannemann, Tobias Senst, Rico Andrich, the members of the Department of Knowledge Processing and Language Engineering, the members of the NIMITEK project consortium, the subjects that participated in the Wizard-of-Oz experiment, and the supporters that took part in evaluating, transcribing and annotating the NIMITEK corpus.

For their love and support, I thank my family.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Although in the last decade we witness the rapid increase of research interest in affected user behavior, it still turns out to be a challenge for developers of spoken dialogue systems. Research in this domain is usually primarily concentrated on detection of emotional user behavior. However, less attention is devoted to another important research question—how to enable dialogue systems to overcome problems in the interaction related to affected user behavior. The latter research question is addressed in this work.

This thesis deals particularly with the research question of adaptive dialogue management in human-machine interaction. To the extent that dialogue management is adaptive, it should take also external factors into consideration when it makes decisions related to the management of the interaction in a broad sense. Examples of external factors include: the emotional state of the user, the context of the interaction, problems that may emerge in the interaction, etc. Adaptivity may be involved at various levels of functionality of the dialogue manager: interpreting user's input, managing contextual information, deciding on the content and the presentation of system's output, etc.

This thesis makes contributions to adaptive dialogue management in the areas of theory, experimental practice, and system development.

1.1 Motivation

Spoken interaction between human and machine just recently became again a burning research issue. Rahwan and McBurney (2007) explain this effect as a consequence of the development of technology that modified our understanding of the nature of computation. They formulate a new metaphor:

computation as interaction, or the joint manipulation of concepts and actions by discrete entities, both human and software agents (Rahwan and McBurney, 2007, p. 21).

One of the widely accepted postulates of human-machine interaction is that it should be as natural as possible. Observing natural language as a database query language, Watt (1968) introduces the term *habitable language* to denote a language in which users can express themselves naturally and without conscious effort to avoid uttering sentences that would not be recognized by the system. Considering advisory systems, Guindon (1988, p. 191–2) introduces a definition of the *habitable natural language interface.* Besides naturalness of the interface language and little effort by the user, Guindon proposes a sufficiently wide and dense language, a small failure rate and robust parsing, informative error messages and fast response time. Similar criteria for user habitability that apply to interfaces are introduced by Carbonell (1986). With their criteria, both Guindon and Carbonell primarily address dialogue phenomena related to the users' language (e.g., syntactically very simple utterances, high frequency of ungrammaticalities, predominant use of ellipsis and anaphora over complex syntactic constructions, meta-language, etc.). They suggest that a habitable language interface should be able to resolve such phenomena. However, they consider also the system's response when a sentence cannot be parsed or understood. The system's response should be *informative enough to allow the linguistically naive user to immediately correct the faulty sentence appropriately* (Guindon, 1988, p. 191) and in a form that is *comprehensible for the user* (Carbonell, 1986, p. 162). Thus, they indicate the need for dialogue strategies that would support the user to overcome problems that may occur in interaction.

Although the above criteria were introduced 20 years ago, the need for dialogue strategies is still present. Besides inherently present dialogue phenomena related to the language, including also affected speech, there are additional reasons that relate to the technology used. Lee (2007) argues that the state-of-the-art automatic speech recognition approaches still cannot deal with flexible, unrestricted language. Bohus and Rudnicky (2008, p. 123–4) state that in settings when systems operate under the conditions of spontaneous speech, large vocabularies and user population, and large variability in input line quality, average word recognition error rates are 20–30%, and they go up to 50% for non-native speakers.

However, inaccurate speech recognition is just one of the reasons that cause miscommunication. In general, problems in the interaction are not caused only by technical deficiencies. It is not reasonable to expect that users will always behave "cooperatively" and that they will produce utterances that fall within the application's domain, scope and grammar. Forcing users to always produce "correct" utterances would significantly limit the naturalness of the interaction. Furthermore, for users in affected states, such a cooperative behavior is hardly to be expected at all. In addition, users may experience problems related to the domain of the application (e.g., the user does not know hot to solve a given task, etc.). Therefore, problems in human-machine interaction appear to be inevitable and cannot be addressed only by careful designs of technical systems.

Adaptive dialogue management is a promising research direction to address the question of handling various problems that may occur in human-machine interaction. The basic functionalities of the adaptive dialogue manager include: modeling contextual information, keeping track of the state of the interaction, and dynamically adapting both analytical and generative aspects of the system's behavior according to the current state of interaction. In other words, this can be formulated as: recognizing that a problem occurred in the interaction, providing support to the user in an appropriate form—tailored to a particular problem and to the user's individual needs—and trying to advance the interaction.

Although the filed of adaptive dialogue management has just begun to evolve, considerable research effort is already to be noticed. We provide an overview of the state-of-the-art approaches in this field, particularly in the following research aspects: collecting emotion corpora (Section 2.3), traditional approaches to dialogue management (Section 3.2), adaptive dialogue systems (Section 4.2), and handling miscommunication in the context of human-machine interaction (Section 5.4.1). However, although adaptive dialogue management is of primary importance for increasing the level of naturalness of human-machine interaction—and, consequently, the level of acceptance of such interfaces by users—its possibilities are by no means sufficiently explored. One of the aims of this thesis is to make a step in this direction.

1.2 Scope and Outline of the Work

This thesis represents a part of the work in the framework of the NIMITEK project[1] (*Neurobiologically inspired, multimodal intention recognition*

[1] The project NIMITEK is funded by grant from federal state of Sachsen-Anhalt, Germany, within the framework of the Excellence Program "Neurowissenschaften". The presented study is performed as a part of the NIMITEK project (FKZ: XN3621A/1005M).

for technical communication systems, cf. Wendemuth et al. 2008) that involves interdisciplinary research on interaction between humans and cognitive technical systems. This interdisciplinary research integrates the fields of computer science, electrical engineering and (neuro-)biology to investigate processing of input, knowledge representation, and decision making in dialogue situations. The project aims at number of scientific goals including: multimodal emotion recognition from acoustic data (i.e., prosodic information), video data (i.e., facial expressions), and textual data (i.e., linguistic features); modeling of the context and adaptive dialogue management; generating *emotionally colored* output; developing (neuro-)biological models of perception, learning and behavior in dialogue situations; etc.

The NIMITEK prototype spoken dialogue system (Figure 1.1) was implemented to demonstrate research achievements in multimodal recognition of emotions and adaptive dialogue management. The central component of this prototype system is the adaptive dialogue management module. At the application level, it was designed and implemented to support users while they solve a task in a graphics system (e.g., the Tower of Hanoi puzzle). More generally, this module illustrates the focal points of adaptive dialogue management presented in this thesis, e.g., interpreting propositional content of the user's commands, modeling contextual information, dynamically adapting the dialogue strategy, providing support, etc.

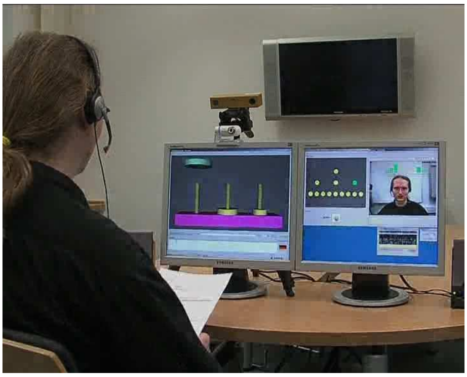


Figure 1.1: The NIMITEK prototype system.

This thesis introduces an approach to adaptive dialogue management in human-machine interaction. We discuss important theoretical considerations and implementation issues in development of an adaptive dialogue management module, and exemplify them for the NIMITEK prototype spoken dialogue system. The introduced approach to adaptive dialogue management is not intended to cover the general case of unrestricted human-machine interaction within arbitrary domains. However, it is not limited to the interaction domain of the NIMITEK prototype system only. With respect to the domain of the interaction, this approach covers the class of spoken dialogue systems that are intended to manage a subclass of task-oriented dialogues, i.e., dialogues that are primarily concentrated on a given task, where the state of the task is observable in the sense that it can be explicitly defined and evaluated regarding to how it corresponds to expected final states. In addition, we concentrate on spoken human-machine interaction in the specific case where some kind of display with a graphical interface is involved. We discuss that display represents an additional non-linguistic context shared between the user and the system, and that it may influence the language of the user (e.g., predominant use of elliptical and minor utterances, context dependent utterances, etc.). With respect to the processing of the user's spoken input of different syntactic forms, the proposed approach covers the class of spoken dialogue systems that are intended to control a subclass of graphical user interfaces, e.g., manipulating with graphical entities represented on the display, controlling graphical menus, solving graphically-based tasks, playing interactive board games that includes spatial reasoning, etc.

The introduced approach to adaptive dialogue management represents an integration of several lines of research: producing and evaluating corpora of affected behavior in human-machine interaction, modeling attentional information on the level of the user's command, and designing adaptive dialogue strategies. In following, we provide an overview of the chapters' content.

Research on emotions in human-machine interaction can be essentially supported by corpora containing samples of emotional expressions. Chapter 2 addresses the question of how to acquire an appropriate emotion corpus. Underlying this question is the problem of assessing phenomenon of affective behavior as it naturally occurs. This problem appears to be the hardest for acquisition of emotion corpora. A fundamental requirement for such corpora is that they have to be ecologically valid, i.e., collected samples should be representative of emotions as they occur in everyday life. The main criticism of existing corpora is leveled against the often used practice of using material produced by actors and disregarding less intense, everyday emotions. As we discuss in this chapter, the essence of this problem lies on the methodological level. Thus, we address the methodological desiderata in obtaining emotion corpora, describe the Wizard-of-Oz experiment conducted in order to produce the NIMITEK corpus of affected behavior in human-machine interaction, and report positive results of the evaluation of the produced corpus with respect to its ecological validity.

The NIMITEK corpus contains 15 hours of audio and video recordings of interaction between the German speaking subjects and the simulated system. All dialogues are transcribed, and dialogue acts are annotated. The corpus had an important role in developing the dialogue management module in the NIMITEK prototype system. Two main lines of research that were supported by the corpus were: modeling attentional information (Chapter 3) and designing an adaptive dialogue strategy for supporting users to overcome problems that may occur in the interaction (Chapter 4).

Chapter 3 proposes an approach to processing of users' commands in human-machine interaction for the restricted model of commands contained in the NIMITEK corpus. Inspection of the NIMITEK corpus showed that the subjects often produced ”irregular” (e.g., elliptical or minor, etc.) utterances. As mentioned above, forcing users to always produce ”well structured” utterances would be too restrictive and not well accepted—especially by users in affected states. Attentional information is already recognized as crucial for processing of utterances in discourse. Thus, we introduce a new model of attentional state—*the focus tree.* We use it to model attentional information on the level of the user's command, and to introduce rules for transition of the focus of attention. The main advantage of this modeling is that, instead of predefining a grammar for accepted commands, we allow more flexible formulation of users' commands. The implementation of this model in the NIMITEK prototype system was demonstrated to work well for different syntactic forms of users' commands: elliptical commands, verbose commands (i.e., the commands that were only partially recognized by the speech recognition module), and context dependent commands. However, the proposed approach to processing users' commands is not limited to commands from the NIMITEK corpus only. We discuss that it is appropriate for the class of spoken dialogue systems that are intended to control a subclass of graphical user interfaces.

Chapter 4 goes further in achieving a higher level of naturalness of the interaction. This chapter proposes an approach to designing adaptive dialogue strategies. More precisely, this chapter reports about design and implementation of the adaptive dialogue strategy in the NIMITEK prototype spoken dialogue system for supporting users while they solve a problem in a graphics system. Again, it should not be understood that this approach is limited to the interaction domain of the NIMITEK prototype system only. We discuss that it covers the class of spoken dialogue systems that are intended to manage a subclass of task-oriented dialogues, where the task states are observable. The main idea is that the system dynamically adapts dialogue strategy according to the current state of interaction. In order to make the system able to select and apply an appropriate adaptation of the dialogue strategy various interaction features should be taken into account. We call the composite of these features *the state of the interaction*. For the purpose of this contribution, we consider five interaction features: the state of the task, the focus of attention (introduced in Chapter 3), the user's command, the state of the user, and the history of interaction. Three requirements underlie dynamical adaptation of the dialogue strategy. First, the user should be provided with useful, sufficient and appropriately emphasized information tailored to a particular problem. Second, support should be timely provided, without relying on the assumption that the user will clearly state a need for support. And third, the manner of providing support should be tailored to meet the user's needs, i.e., it should be in accordance with the emotional state of the user. Dynamical adaptation of the introduced dialogue strategy includes three distinct but interrelated decision making processes that reflect these requirements: When to provide support to the user? What kind of support to provide? How to provide support? These decision making processes are considered in more detail. Finally, we provide a brief overview of the functionality of the dialogue management module and of its relations to functionalities of other modules incorporated in the NIMITEK prototype system.

While these chapters introduce and illustrate various theoretical considerations and implementation issues related to adaptive dialogue management, several important research questions remain to be discussed. Some of them are: How do the proposed algorithms work in a realistic scenario? To what extent is the proposed modeling approach task-independent? Can we extend the model of attentional state to cover more interaction domains? Can we extend the dialogue strategy in order to provide long-term support to the user? How could the introduced models be used to handle miscommunication in the context of spoken human-machine interaction? Chapter 5 addresses these questions. First, we analyze an actual dialogue between the user and the NIMITEK prototype system that took place during the testing of the system. Second, we introduce and discuss the implementation of an extension of the adaptive dialogue strategy aimed to provide long-term support to the user. Finally, we discuss and illustrate how the dialogue management module handles miscommunication on different levels: the conversational level, the intentional level, and the signal level.

Chapter 6, the final chapter of the thesis, contains conclusions and future prospects of research. In Appendix A, we employ the NIMITEK corpus as a tool that provides an empirical foundation for analyzes of emotional content based on linguistic information derived from the transcribed dialogues. We shortly discuss various linguistic features that may carry affect information (e.g., key words and phrases, lexical cohesive agencies, dialogue act sequences, etc.). Appendix B provides a description of the set of graphically-based tasks that was given to the subjects in the Wizard-of-Oz experiment reported in Chapter 2.

Chapter 2

The NIMITEK Corpus

2.1 Introduction

It is a widely accepted fact that research on emotions in human-machine interaction can be essentially supported by corpora containing samples of emotional expressions. It is thus not a surprise that an emotion corpus—the NIMITEK corpus—had a crucial role in the development of the spoken dialogue system described in this thesis. The coming chapters report how various insights that resulted from analyzes of this corpus were integrated in the conceptual design and implementation of the dialogue management module in the NIMITEK prototype system.

There are many important research questions that are to be properly addressed when implementing an emotion-aware dialogue system. One of them raises at the very beginning: How to acquire an appropriate emotion corpus? Underlying this question is the problem of assessing the phenomenon of affective behavior as it naturally occurs. This problem appears to be the hardest for acquisition of emotion corpora. As we discuss below, the essence of this problem lies on the methodological level. One of the aims of this chapter is to address the methodological desiderata in obtaining emotion corpora. The chapter can be summarized in the following points:

- First, we propose requirements that are to be met in order that a Wizard-of-Oz (WOZ) scenario designed to elicit affected behavior could result in ecologically valid data.
- Second, we describe the WOZ experiment conducted in order to collect the NIMITEK corpus of affected behavior in human-machine interaction.

- And finally, we report results of the evaluation of the NIMITEK corpus with respect to its emotional content and we demonstrate a satisfying level of its ecological validity.

At the end of the chapter, we briefly indicate two main lines of research represented in this thesis that were supported by the NIMITEK corpus.

2.2 The Notion of Ecological Validity

Douglas-Cowie et al. (2004) define the kind of corpus that is needed to support the development of emotion-sensitive interfaces, and assess what has been achieved in the field. They emphasize the *naturalism* of emotional content as one of fundamental requirements—collected samples should be representative of emotions as they occur in everyday life (Douglas-Cowie et al. 2004, p. 7). The notion of naturalism is closely connected to the notion of *ecological validity.* Brewer (2000, p. 3) explains it as a form of generalizing from the obtained results that is based on the question of *whether the effect is representative of what happens in everyday life.* He notes that ecological validity may be essential when the research agenda is descriptive (Brewer, 2000, p. 12). By the term *descriptive research*, Brewer (2000, p. 3) primarily refers to *research undertaken for the purpose of demonstration [...] conducted in order to establish empirically the existence of a phenomenon or relationship.*

Although the significance of ecologically valid corpora seems to be widely recognized, Douglas-Cowie et al. (2004) conclude that this requirement is not adequately addressed in existing corpora. They level their criticism against the often used practice of:

- using material produced by actors:

> [...] many databases that are described as having emotional content are in fact acted or posed [...] These have played a key role in developments to date, but there is evidence to show that data of this type does not bear a straightforward relationship to emotion in everyday life. (Douglas-Cowie et al. 2004, p. 6)

- disregarding less intense emotions present in everyday life:

> The data should also reflect the type of emotion that occurs in everyday life. Existing databases are often oriented

> to extreme representations of a few emotions (often based on traditional lists of 'primaries'), and although these do indeed occur in everyday life, much of our life is taken up with less intense emotion (irritation rather than anger, pleasure rather than elation) or emotion-related states (such as friendliness, interest, satisfaction, stress, anxiety). (Douglas-Cowie et al. 2004, p. 7)

The essence of this criticism lies on the methodological level. The methodologies used so far to collect emotion corpora can be classified in the following groups (Douglas-Cowie et al. 2004, p. 15-6):

1. collecting acted data, e.g., asking actors to simulate emotions according to a predefined scenario;

2. collecting application-driven data, e.g., recording telephone conversations in call-centers, etc.;

3. collecting induced data, i.e., stimulating subjects to produce an emotional response in laboratory conditions;

4. collecting naturalistic data, i.e., collecting real-life, everyday samples in *field* settings.

These methodologies vary a great deal with respect to the artificiality of settings, with no agreement between researchers on the most productive method. On the one side of the scale is collecting emotional expressions from professional and non-professional actors; on the other side are attempts to record "real life" situations. The first approach is strongly criticized for the reason that samples collected in this way are not representative of everyday emotions; the second approach is related to many nearly unsolvable problems of controlling the environment.

The question that remained open is how to collect emotion corpora. To illustrate this, let us make a simplification of our long-term research aim—we want to develop a spoken natural language dialogue system that should be able to perform two tasks: (1) to determine, based on the recognition of negative user's emotional states, critical phases in human-machine interaction and (2) to resolve the problem that emerged in interaction by applying an appropriate dialogue strategy. Brewer (2000, p. 3) emphasizes that we cannot speak of the validity or invalidity of research per se—*validity must be evaluated in light of the purpose for which the research was undertaken* *in the first place.* Thus, for the former demand, it is preferable that the material contained in the used corpus was collected from people experiencing genuine emotions rather than produced by actors. For example, classification results of a statistical prosodic classifier for emotion recognition from user's spoken input may depend to a high extent on the fact whether it was trained on genuine or acted expressions of emotions (cf. Batliner et al. 2000). Eliciting emotions in laboratory setting is a challenging task; therefore a research in *field* settings (e.g., collecting real-life samples) might be preferred for the purpose of assessing the phenomenon of affected speech as it naturally occurs. However, such a decision is not in accordance with the latter demand.

To satisfy the latter demand, it would be useful if the underlying corpus contained samples of dialogues between the user and the system that provide insight in various dialogue strategies that could be applied in order to resolve problems emerged in interaction. In order to produce such samples, researchers should have the possibility to control development of the dialogue between the subjects and the system during the derivation of samples. However, this is usually possible only in laboratory settings.

Thus, the question of choosing an appropriate acquisition methodology, restricted in its scope, can be reformulated: how to satisfy these methodological requirements that appear to be contradictive. In general, this question is still not addressed properly. In Section 2.3 we discuss briefly the state-of-the-art corpora with realistic emotions. Then we go a step further—we address the methodological desiderata in Section 2.4.

2.3 Background and Related Work

We do not aim here to provide a complete list of all corpora with realistic emotions (an overview is available via the HUMAINE portal, cf. Douglas-Cowie et al. 2004), but rather to highlight some of them that we find relevant for our discussion. One of the most significant corpora with genuine emotions is the HUMAINE Database (Douglas-Cowie et al., 2007). One of the key goals of the HUMAINE project is to provide the community with examples of the diverse data types that are potentially relevant to affective computing. Work on the HUMAINE Database began in response to the observation that the requirement of ecological validity was not adequately addressed in existing corpora (Douglas-Cowie et al., 2007, p. 488-9). This database was designed to provide a concrete illustration of key principles. It is a collection of different emotion corpora, and the recorded material can be classified in two categories: naturalistic data and induced data.

Naturalistic data in the HUMAINE Database. The following three databases containing naturalistic data were mainly collected from different TV programs:

- *Belfast Naturalistic Database* (Douglas-Cowie et al., 2000) contains recordings from two sources. The first source was recordings of discussions between *people who knew each other well talking about emotive issues* (Douglas-Cowie et al., 2000, p. 41). The second source was recording of audiovisual sedentary interactions from TV chat shows and religious programs.

- *EmoTV Database* (Devillers et al., 2006; Devillers and Martin, 2008) contains recordings of audiovisual interactions from TV interviews. This corpus is focused on *spontaneous expression of emotion during monologues* (Devillers and Martin, 2008) in which interviewed people talk on issues such as lawsuit, election campaign, problems of environment, increase of the price of coffee, football cup, etc.

- *Castaway Reality Television Database* (cf. Douglas-Cowie et al. 2007, p. 490) contains audiovisual recordings of people taking part in a competing activities on a remote island (e.g., feeling snakes, lighting outdoor fires, etc.). The recordings include also post-activity interviews with participants.

With respect to the emotional content of these corpora, Douglas-Cowie et al. (2007) note:

> All of these were chosen to show a range of positive and negative emotions. Intensity is mostly moderate, though EmoTV and Castaway contain more intense material (Douglas-Cowie et al., 2007, p. 490).

However, even though such TV programs could provide emotional material, the drawbacks of this approach are the lack of control over the environment, especially over dialogue development, and the lack of repeatability of the scenario. An alternative is to collect recordings that contain expressions of emotions induced in laboratory settings. This leads us to the second category of material in the HUMAINE Database—induced data.

Induced data in the HUMAINE Database. Several different techniques for induction of emotions were used in collecting the HUMAINE

Database. Two of them are particularly important for this discussion: *The Sensitive Artificial Listener* (SAL) and *The EmoTaboo Protocol.*

The Sensitive Artificial Listener (Douglas-Cowie et al., 2008) is an induction technique intended to generate data that are both suitable for machine analysis and reasonably natural. It was inspired by a technique for inducing emotions that is often (successfully) applied in talk-shows. Moderators in such shows usually do not involve themselves much in dialogue. They just maintain conversation by uttering short questions and statements in order to additionally increase the level of emotionality in the conversation on a topic that a guest already perceives as emotional. In the SAL technique, the focus is on conversation between a human and an agent that either is or appears to be a machine. Responses of the system are adjusted to emotional coloring of what the subject says. The human operator playing the role of the system simulates one of four characters that aim to make people *happy*, *gloomy*, *angry* and *pragmatic*, respectively (cf. Douglas-Cowie et al. 2008, p. 1). Subjects can choose at any time to which character they want to talk. The response of the operator depends on the character that is active and the user's state. For each character there is a set of predefined responses encouraging the user into responding in different emotional states. Different versions of SAL are developed in English, Hebrew and Greek. Positive results are reported:

- The SAL scenario has been used successfully in three major EU projects (ERMIS, HUMAINE and SEMAINE) to generate large amounts of data that has been labeled and used in a machine learning context (Douglas-Cowie et al., 2008, p. 2). Four raters have participated in the evaluation of the data with respect to the emotional content. They have labeled the data using the FEELTRACE tool (Douglas-Cowie et al., 2000, p. 43-4). It uses two dimensions that are commonly related to emotion—arousal and valence—in order to describe an emotional state in terms of continuous scales. Arousal-valence space is represented by a circle on a computer screen, and raters describe perceived emotional state by moving a pointer to the appropriate point in the circle using a mouse. In addition, using the FEELTRACE tool, it is possible to track emotional content over time and to explore the relationship between time and modality in the expression of emotion.

- The data generated is rich in facial and non verbal signals (e.g., aspects of pitch, spectral characteristics, timing), and shows a considerable range of emotions and emotional intensities (Douglas-Cowie et al., 2008, p. 3).

- The SAL data that is already available is of sufficient quantity and quality to train machine recognition systems (Douglas-Cowie et al., 2008, p. 4).

We make two comments related to this technique. First, in the SAL technique, the starting point for induction of emotions is a dialogue topic that a subject perceives as emotive. The technique does not manipulate subjects' emotions. It rather *gives them prompts to which they can react emotionally if they choose to* (Douglas-Cowie et al., 2008, p. 2). The authors note that using this technique it is easy to build up quite a high level of involvement during a sequence of exchanges on such *an emotive topic*. However, it is still an open question whether this technique can be successfully applied for dialogue topics that a subject does not *a priori* perceive as emotive. Second, this technique does not try to control development of the dialogue. The SAL has no intelligence—the responses of the system are predefined, and scripts followed by the operator do not take the context of the interaction into account. As we discuss in the next section, applying flexible dialogue strategies, instead of only fixed patterns of messages, might be valuable for investigating different possibilities of dialogue management.

The EmoTaboo Protocol (Devillers and Martin, 2008) is an adaptation of the game Taboo. In this game, one of the players has to guess a word that the other player is describing using speech and gestures, but without uttering five forbidden words. The word to guess and the five forbidden words are written on a card. The EmoTaboo Protocol involves interactions between two players that do not know each other: a naïve subject and an instructed player. Strategies for eliciting emotions from the naïve player were used at three different levels:

- The course of the game: The players were given limited time to read a card and to make their guess. If the secret word was not found in the given time or if the players violated the rules of the game, they received penalties.

- The selection of the cards: Cards were provided to the players in ascending order of difficulty. They contained very uncommon words (e.g., "palimpsest") supposed to arouse embarrassment or shame, words evoking disgusting things (e.g., "putrid") or words with sexual connotation (e.g., "aphrodisiac").

- The instructed player: For each card, the instructed player received a set of emotions to elicit from the naïve player. The instructed player used different strategies to elicit emotions: e.g., intentionally proposing words with no relation to what naïve player said, criticizing the naïve player, etc.

A range of emotions, including embarrassment and amusement, is reported.

We note three aspects of this technique. First, in contrast to the SAL, this technique is aimed to manipulate subjects' emotions. In addition, involving the instructed player gives the possibility to have some control over the dialogue development. Second, the experimenters tried to address the problem of the role-playing subjects. To increase the engagement of the subjects in the given task, a gift token was promised to the winner team. And third, this technique is focused on human-human interaction.

Both these induction techniques illustrate aspects of an acquisition methodology appropriate to satisfy demands both for genuine emotions and for control over dialogue development. In our approach, we apply also an induction technique in order to collect an emotion corpus. Our focus on human-machine interaction allows us to use the Wizard-of-Oz (WOZ) technique. The underlying idea is that subjects are given the illusion that they are communicating with a computer system, while a human operator plays the role of the system. In the next section, we discuss the WOZ technique in detail. We consider corpora produced using the WOZ technique and discuss different implementations of this technique. Finally, we introduce a refinement of the WOZ technique that was used to produce the NIMITEK corpus.

2.4 Refinement of the WOZ Technique

As already mentioned, the concept of ecological validity of a study relates to the extent to which laboratory settings reflect "real life" conditions. Stressing the importance of ecological validity of corpora used in research on the expression of emotions, Douglas-Cowie et al. (2000, p. 39–40) introduce four guidelines to achieve it:

1. genuine emotion—using material generated by people experiencing genuine emotions, rather than material produced by actors,

2. emotion in interaction—focus on examples derived from people engaged in interactions,

3. gradation—sampling situations where emotions are mixed or controlled in the ways that typically occur in everyday life,

4. richness—collecting samples that are extended both in modality (vocal expressions, facial expressions, gross gestures, etc.) and time.

We mentioned above that we apply the WOZ technique in order to produce the NIMITEK corpus. Fraser and Gilbert (1991, p. 82–83) consider three basic requirements that have to be met for a useful WOZ simulation: the simulation must be possible given human limitations; the future system must be specifiable; and the simulation must be convincing. Nevertheless, besides the obvious advantages of this technique over recording emotional expressions produced by actors, even the fulfilment of all commonly established requirements for a useful WOZ simulation does not guarantee that subjects will experience and signal genuine emotions. We propose additional requirements that are to be met in order that a WOZ scenario designed to elicit affected speech in human-machine interaction could result in ecologically valid data.

Generally speaking, in order that any tactic used for the emotion elicitation in WOZ scenarios is to be effective, two additional requirements are to be met. First, subjects have to be motivated to accomplish a given task in order that a successful accomplishment or a failure to accomplish could induce an emotional state. Even then, the factual induction of an emotion does not itself constitute its manifestation. Therefore, the second requirement is that subjects have to be stimulated to express themselves, so that their induced emotions can be signaled overtly.

We provide a justification for these requirements by observing already conducted WOZ experiments primarily related to prosodic properties of affected speech. This should not be understood as a criticism of these experiments. We see them rather as valuable experiences that led us to the conclusions presented in this chapter.

2.4.1 Addressing the Role-Playing Subjects

The first requirement is introduced to address the problem of role-playing subjects. A point of departure for this critique against the artificiality of experimental settings is that subjects in WOZ experiments are aware of laboratory settings, so they do not behave as real users. They are usually not really motivated to accomplish given tasks and it is thus rather difficult to induce emotions in them. Experimenters are generally aware of these laboratory effects, even though it is not always clear to what extent these effects influence results. Still, the level of motivation of subjects to accomplish a given task is often considered to be a confounding factor in WOZ experiments, i.e., the factor over which the experimenter has no control. Pirker and Loderer (1999) report a WOZ simulation of a speech based ticket reservation system, and present the findings on the prosodic properties of repeats and corrections triggered by intentional rejections or misrecognitions of subjects' utterances. In concluding, they note:

> [...] it has to be kept in mind that all users of the system were basically role playing. They are no real users with real information requirements, real time constraints or even real telephone bills. Nevertheless this laboratory effect probably only has a minor impact on the aspects of prosody described in this paper. (Pirker and Loderer, 1999, p. 185)

Although they are aware of limited motivation of subjects participating in the experiment, their remark about the minor impact on the aspects of prosody is not of a general nature. In certain WOZ experiments, especially in those considering the emotional behavior of subjects, this issue turns out to be crucially important. A similar tactic for emotion elicitation, i.e., the simulation of malfunctioning of a system for translation of spontaneous speech, was used in the WOZ experiment conducted in the framework of the Verbmobil project. The scenario was designed *to provoke reactions to probable malfunctions and to control the speakers changes in attitude towards the system, i.e. their emotional behavior, over time* (Batliner et al., 2000, p. 196). However, data collected in this WOZ experiment were not the only data used for the training of statistical classifiers. In addition, an experienced acting person was asked to read neutral sentences from the Verbmobil scenario, and to imagine the situations in which the system was malfunctioning in order to produce sentences using angry emotional prosody. Also, a group of naive subjects was asked to read prefabricated sentences using neutral and angry emotional prosodies. By contrast to the former mentioned experiment, a major impact of role-playing subjects on the experimental results is reported—the classification results for the WOZ data were the worst. Batliner et al. observe that subjects in the WOZ experiment used prosody less and did not necessarily signal their emotions overtly. As possible reasons they state that actors displayed emotions overtly because they have been asked to do so, and that in the read aloud scenario the use of prosody is the only strategy available, while subjects in the WOZ experiment may use different communicative strategies besides the use of prosody (Batliner et al., 2000, p. 199). Even so, discussing the concept of WOZ scenario in general, they note:

> Still, it is 'as if' again since even if the subjects do believe that they are communicating with a real computer, they just pretend to need some information; normally, they are very co-operative, and that means that it is rather difficult to make them really angry. At least, one can never be sure that they would behave the same way in a real life task. (Batliner et al., 2000, p. 196)

In another WOZ experiment, conducted to collect a corpus of emotional children's speech, Batliner et al. (2004) partially address the problem of role-playing subjects. In this experiment, the children were asked to guide a dog-like Sony's AIBO robot around a map that was printed on a floor carpet. In the scenario designed to elicit emotional behavior, the children believed that the robot was reacting to their spoken instructions, while the robot, controlled by a human operator, strictly followed a predefined sequence of actions. The children's speech is intended to be *'natural' because children do not disguise their emotions to the same extent as adults do* (Batliner et al., 2004, p. 171). In addition, it was also intended to be *spontaneous, because the children were not told to use specific instructions but to talk to the AIBO like they would talk to a friend* (Batliner et al., 2004, p. 171). Discussing results of the analysis of the corpus collected in this experiment, Batliner et al. (2008, 2006, 2005) report, inter alia, positive results related to its ecological validity:

> The emotions and emotion-related states expressed by the children are 'realistic' in the above mentioned sense: they are not only acting 'as if' they were giving commands. (Batliner et al., 2008, p. 182)

The methodological approach in this WOZ experiment is based on the (reasonable) assumption that children do not disguise their emotions. Still, it addresses the problem of the role-playing subjects only partially—it is demonstrated to work fine for children, but it is clear that such a scenario is generally not appropriate for adult subjects. Nevertheless, an experiment lying out of the WOZ framework, conducted by Aharonson and Amir (2006), goes a step further into the direction of emotion elicitation in subjects while they are speaking. They adapted Damasio's card game (Damasio, 1994, p. 212–217) that has been shown—by measuring skin conductivity of subjects and observing their affected behavior and physiological responses—to elicit apprehension. In their computerized gambling game the four doors displayed on the screen were opened in response to voice commands from the subjects.

By opening a door subjects gained or lost an amount of points. The subjects were told that they could achieve the highest gain if they figure out which set of doors to choose and in which order. Aharonson and Amir (2006, p. 180) report that *subjects were fully concentrated on the game and paid little attention to the lab environment.*

At this point, it seems to be intuitively clear that the application simulated in a WOZ experiment could serve as possible motivation catalyst. Fraser and Gilbert (1991, p. 91) note that WOZ simulations can vary in the basic task they accomplish, but there is only a very limited number of basic tasks. The basic tasks include database querying and updating, modality translation, (i.e., the kind of activity required in "listening typewriter" tasks), and dialogue management. However, there are many possible applications, allowing experimenters to adapt their scenarios.

Forbes-Riley et al. (2008a) present the *Uncertainty Corpus* that contains spoken dialogues between students and WOZ spoken dialogue tutoring system. Considering student affects, they target uncertainty. Their starting point was that uncertainty is inherently present in tutoring dialogues as a signal of "learning impasses". In the experimental settings, the system poses a conceptual physics problem. For each student answer, the wizard performs speech recognition, correctness annotation, and uncertainty annotation. Upon hearing a student answer, the wizard uses the experiment interface software to annotate whether the answer was correct or uncertain, respectively (e.g., an answer may be correct and uncertain in the same time). Forbes-Riley et al. report that uncertainty occurs more often then other student affective states in their dialogues—796 out of the 2171 dialogue turns produced by students were annotated as uncertain.

Another interesting methodology for collecting spontaneous expressive speech that relies on a WOZ scenario-based induction of affective states is introduced by Fék et al. (2008). Their WOZ experiment was set up using the dedicated software platform E-Wiz. The subjects were asked to test a futuristic phonetic-learning tool that enables users to easily produce vowels of foreign languages using their brain plasticity. In order to induce emotions, the tool was manipulated by the wizard: To induce positive emotional states, subjects' performances were evaluated in the first training phase as better than previous mean performances, and in the second phase as within the best three performances ever realized, allowing them to skip directly to the last and more complex phase. Negative emotional states were then induced by evaluating subjects' performances as very bad and telling them that their perception skills might have been damaged by the software. This methodology was used to collect corpora in French and Hungarian. A perceptual test was performed to assess the emotions expressed in the collected speech data, and to retain a set of collected expressions for which the self-evaluation—with respect to emotional content—provided by the subject matched the judgement of naïve listeners. In the first phase of the perception test, 6 French and 7 Hungarian subjects were asked to self-evaluate expressions of the affects they had just been feeling. These self-reports were classified by the experimenters into four categories: {anger, fear, joy, other} for the French corpus, and {satisfaction, dislike, stress, other} for the Hungarian corpus. In the second phase of the perception test, recordings containing emotional expressions produced by the French and Hungarian subjects during the experiment was additionally evaluated by 15 French and 25 Hungarian naïve listeners, respectively. They were allowed to listen to the selected recordings as many times as they wanted and had to select one of the predefined emotion classes. Fék et al. report that the validation of the self-evaluation was successful for 79 of the 146 spontaneous speech expressions in the French experiment, and for 52 of the 103 expressions in the Hungarian experiment.

2.4.2 Controlling the Dialogue Development

The second requirement for successful emotion elicitation in WOZ experiments is that subjects are stimulated to express their emotions overtly. It should be clear that this is by no means a logical implication of the first requirement. To illustrate this, let us go back to the experiment conducted by Aharonson and Amir (2006). As mentioned above, the design of their computer controlled environment is based on a gambling game that was proven to be motivating for subjects. But to assess the usability of obtained data, the overall experimental settings should be taken into consideration. The subjects in this experiment were instructed to use only two sentences ”open this door” and ”close door”, while they indicated a door to open with a mouse. The intention of experimenters was to control the textual and emotional content, i.e., to collect a number of identical, short phrases uttered with different emotional prosodies and to link them with the events of gain or loss of the points in the game.

Besides the positive aspects of minimizing the subjects' awareness of the laboratory settings, the described experimental settings suffer from two disadvantages. First, the obtained corpus consists of a limited number of short phrases from a small vocabulary only. We recall that researchers in this experiment, as well as in all so far mentioned WOZ simulations, were interested in prosodic properties of affected speech. The suprasegmental nature of prosody is illustrated by Martin (1996) in a convincing manner. Considering prosodic realization at both clause and text levels, he demonstrates that prosodic structures map over a range of segments (Martin, 1996, p. 42–43, 50–51)[1]. Thus, collecting only short samples of emotional expressions might not suffice for all analyzes of affective prosody, because they cannot reflect the prosodic realization on higher levels in the linguistic system. In other words, the collected samples of emotional expressions should be extended in time. This leads us to another point—the significance of dialogue control in WOZ simulations.

The second disadvantage of the described settings is leveled against the computer environment; namely, it does not take the human-machine interaction into account. A similar observation holds for the above mentioned WOZ experiment with a dog-like Sony's AIBO robot. This experiment was primarily concentrated on collecting emotional speech, without taking interaction into account. The robot did not produce verbal output, so no spoken dialogue was collected. Also, it followed predefined sequence of actions, so no flexible system's interaction strategies were applied. Automatic determination of critical phases in human-machine interaction is an important task; however, it is only a half of the solution. The other half is to resolve the problem emerged in interaction. It is not surprising that problems of spoken dialogue systems do not relate only to speech recognition, but also to dialogue management. Existing restricted system-guided dialogues are not well accepted by most users, and a more sophisticated approach to dialogue adaptation is needed. Therefore, the experimental settings should allow experimenters to observe and control the development of the dialogue between subjects and the simulated system. This requirement influences also the wizard's behavior. For the purpose of investigating different possibilities of dialogue management, it is an advantage if the wizard uses more flexible dialogue strategies, instead of only fixed patterns of messages. However, considering the dialogue structure and content, we should also mention the importance of wizard's response time.

2.4.3 Wizard's Response Time

As one of the reasons why the wizard's response time is important, Fraser and Gilbert (1991, p. 94) mention that *speed of response can be expected*

[1] Especially indicative is Martin's choice of the parameter *affect* to illustrate prosodic text structure—the parameter *deployed to negotiate solidarity with the listener/reader* (1996, p. 51). Addressing the role of affective prosody, he additionally underlines Halliday's (1994) assertion that interpersonal meaning uses prosodic principles.

to affect dialogue structure and content. Guindon (1988, p. 192) proposes that in a habitable natural language interface the response to the user's utterances should be fast enough not to interfere with performance of the main task. He interprets the low frequency of pronouns in human-machines dialogues as one of the evidences that subjects believe that there is *a poor shared (linguistic) context* between them and the system. However, he is aware that the slowness of the interaction between subject and wizard may also explain the low frequency of pronouns:

> The slowness of typing requests for help and in receiving the answer [...] probably produces a context shift with every request. When context shifts, the entities in focus (which can be referred to by using a pronoun) get out of focus and it becomes necessary to use a non-pronominal noun phrase to refer to them. (Guindon, 1988, p. 194).

Similarly, in the WOZ simulation presented by Pirker and Loderer (1999) utterances were produced with a very slow rate of speech. They observe that many subjects not only failed to omit a redundant noun or at least to deaccent it answering the question of the wizard but often also overarticulated the whole word. For example, when asked by the system "How many tickets do you want to reserve?" users answered with "Two tickets", where the word "ticket" in the answer was not only redundant but also often prosodically marked. They conclude that *this might be due to the effect that unnaturally long pauses within the dialogue may block the linkage to prior mentioned items.* In addition, they note that *it is not clear whether the decreasing speaking rate of some users was due to an adaption to the slow speed of the synthesized utterance* (Pirker and Loderer, 1999, p. 184–185). Therefore, the wizard's response time should be fast enough not to influence speech performance of subjects.

2.4.4 Pitfalls of Proposed Requirements

We proposed two requirements that are to be met in order that a WOZ scenario designed to elicit affected behavior could result in ecologically valid data. This subsection briefly states pitfalls of the proposed requirements.

The first requirement is that subjects have to be motivated to accomplish a given task in order that a successful accomplishment or a failure to accomplish could induce an emotional state. We discussed that the application simulated in a WOZ experiment could serve as possible motivation catalyst. A pitfall of this requirement is related to the question whether ecological validity of obtained data suffices. The WOZ technique is introduced in the first place to address the problems of designing and evaluating speech-based dialogue systems. The simulation is often designed keeping in mind the requirements of a prospective application. Any variation of the simulated application in order to make it more motivating for subjects raises the question of usability of obtained data. If the simulated application differs essentially from the prospective application, the obtained data, though ecologically valid with respect to emotional content, might not be useful for the purpose of designing the prospective application.

The second requirement for successful emotion elicitation is that subjects have to be stimulated to express themselves, so that their induced emotions can be signaled overtly. To achieve this, we proposed that the experimental settings should allow experimenters to observe and control the development of the dialogue between subjects and the simulated system, and the wizard should use more flexible dialogue strategies (including strategies intended to provoke emotional reactions in subjects) instead of only predefined scripts. There is also a pitfall related to this requirement. A wizard may use flexible dialogue strategies, but only to a certain level. The illusion that the wizard is a computer must never be destroyed.

The WOZ simulation described in the next section illustrates how to meet the proposed requirements, and how—simultaneously—to avoid the mentioned pitfalls.

2.5 Collecting the NIMITEK Corpus

This section reports the WOZ experiment[2] conducted in the framework of the NIMITEK project. The simplified schema of the laboratory settings is given in Figure 2.1. The experiment was conducted in two rooms—the subject's room and the wizard's room. The PC in the subject's room represents the simulated spoken natural language dialogue system. The keyboard and the mouse are removed, so subjects can interact with the system only verbally. The wizard performs speech recognition, remotely controls the interface of the system, and produces verbal output of the system. The image of the desktop of the subject's PC and the subject's video are displayed on two separate monitors in the wizard's room. Our experiment deals with expressions of emotion in two modalities at a time: vocal and facial expression. The fixed positioned digital camera in the subject's room captures the

[2] The experiment was approved by the ethics committee of the University of Magdeburg.

face of the subject. The microphone in the subject's room captures verbal interaction between the subject and the simulated system. The joint audio/video signal is saved on the hard disk of the wizard's recording PC. In addition, the desktop of the subject's PC was also recorded. It is, therefore, possible not only to study correlation between vocal and facial expressions, but also correlation between linguistic and non-linguistic contexts shared between the subject and the simulated system.

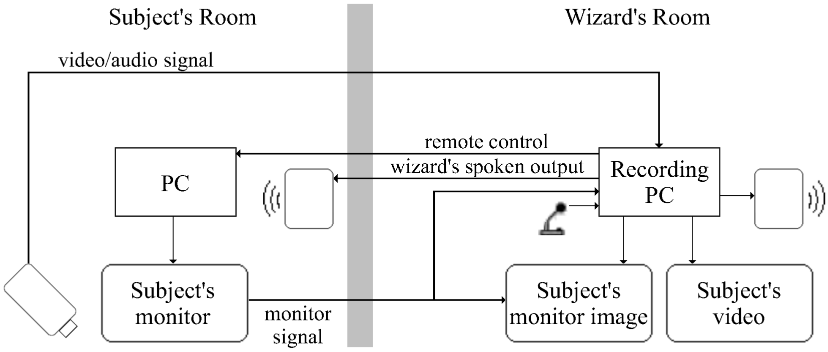


Figure 2.1: Simplified schema of WOZ laboratory settings.

In this WOZ experiment we used a hybrid approach to emotion elicitation—a combination of a motivating simulation environment and different strategies of the wizard. The application planned for the WOZ simulation was a spoken natural language dialogue system that supports subjects while they use a graphical tool. This graphical tool should have been implemented using an existing software platform[3] that allows visual representation and manipulation of graphical objects. It was, however, obvious that it would be almost impossible to motivate normal subjects enough to experience emotions while using this application. Thus, we modified the application in order to make it more motivating for subjects. We decided to simulate an intelligence test because it was expected to be a strong motivational factor for subjects. At the same time, the attractiveness of the simulation environment was preserved and the task variable was not changed—the test was implemented using the existing graphical software platform.

Subjects in our WOZ experiment were asked to undertake a combined test of both intelligence and communication abilities supported by the spoken natural language dialogue system. The test consists of 14 tasks that can

[3] This software platform was developed at the Fraunhofer Institute for Factory Operation and Automation IFF, Magdeburg, Germany.

be classified in 6 groups:

- Filling empty place (3 tasks)—The subject should select one of the four given pictures that logically continues an existing array of pictures.
- Classification (2 tasks)—A 3D-figure and a group of 2D-nets are presented to the subject. For each 2D-net the subject should say whether it represents the given 3D-figure unfolded in 2D or not.
- The Tangram puzzle (3 tasks)—The puzzle consists of seven Tangram 2D-objects (e.g., triangles, etc.). The goal of the puzzle is to form a specific shape using all seven objects.
- The Grid puzzle (3 tasks)—The puzzle consists of an [n x n] grid that contains $n^2 - 1$ tiles numbered from 1 to $n^2 - 1$, and the n^2th place is empty. The tiles are scrambled and the objective of the puzzle is to unscramble the tiles to get them into consecutive order. The subject is allowed only to make moves which slide tiles into the empty space.
- The Tower of Hanoi puzzle (2 tasks)—The puzzle consists of three pegs and several disks of different sizes. The disks are stacked in order of size on the leftmost peg. The goal of the puzzle is to move the entire stack to the rightmost peg according to the following rules: only one disk can be moved at a time, each move consists of taking the upper disk from one of the pegs and placing it onto another peg, and no disk may be placed on top of a smaller disk.
- The Three Jugs Problem (1 task)—There is an eight liter jug of water, and two empty jugs—a three liter jug, and a five liter jug. The objective of this task is to measure exactly four liter of water. Subject is allowed only to pour water from one jug to another jug.

A more detailed description of the tasks is given in Appendix B.

The tasks were successively displayed on the screen with accompanying descriptions spoken by the system. In order to force subjects to verbally interact with the system, they were only allowed to produce spoken instructions to the system (e.g., what operation to perform or to ask the system for a help, etc.), while the keyboard and the mouse were removed. The test was so specified that subjects might use a limited number of different words to solve a task. On the other hand, they had to produce a number of utterances to accomplish the whole test. Only the following subset of three predefined instructions accepted by the system (”Start the test”, ”Task completed”, ”I give up. Next task.") was explicitly given to subjects. Guessing and formulation of other acceptable instructions and questions were introduced as a part of the test as an additional stimulus for subjects to express themselves verbally. In this respect, subjects were given freedom of expression.

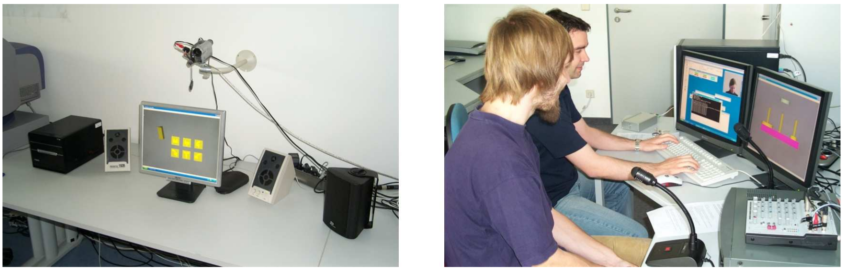


Figure 2.2: WOZ scenario. (Left) Subject's side: Tasks are displayed on the screen with accompanying descriptions spoken by the system. (Right) Wizards' side: one wizard controls the test interface, the other produces utterances.

According to Fraser and Gilbert (1991, p. 93), one of the hardest tasks for the wizard is to restrict his speech recognition capabilities to variables defining the ranges of acoustic, lexical, syntactic and pragmatic phenomena which he is allowed to recognize. Here, the wizard was constrained as to understanding and utterances production only to a level not to destroy the illusion that he is a computer. Additional attention was devoted to the wizard's response time. Therefore, in our experiment we use a cooperating team of two wizards: one producing the utterances, and the other controlling the test interface. In this thesis they are occasionally referred to in singular. Wizards understood utterances that contained anaphora and ellipses, as well as complex utterances.

Ten healthy native German speakers (7 female, 3 male) in the age from 18 to 27 (mean 21.7) participated in the experiment. None of them had educational background or user experience related to state-of-the-art spoken dialogue systems. They gave written consent that their voices and facial expressions may be recorded.

The language used in the experiment was German. The average duration of an experimental session was approximately 90 minutes. During that time subjects were given two pauses of several minutes.

The first, short part of the session was devoted to build subject's trust in the performance of the simulated system. The wizards correctly recognized subject's utterances, performed requested operations as they were instructed, and provided useful comments and answers to subject's questions. In addition, the first group of the tasks given in the test was relatively uncomplicated to solve, which—in combination with good performance of the system—was planned to give the subject an impression that she makes good progress in solving the test. The subject motivated in this way became fully concentrated on the given tasks and they paid only little attention to the lab environment (e.g., the camera, etc.).

In the second, significantly longer part of the session, the wizards simulated malfunctions of the system with intention to provoke emotional reactions of the subject. The premise for emotion induction was that the subject is motivated to achieve a good result in the intelligence test and that she believes to know how to solve a given task. Confronting the subject with simulated malfunctions of the system (e.g., inaccurate speech recognition, etc.) that prevent her from successful solving of the given task was expected to induce emotional states. The wizards simulated the following malfunctions in order to induce negative emotional states in the subject:

- Deliberate misunderstanding of subject's commands and performing incorrect operations that draw back the state of the task from the expected final state. This was shown to be an especially effective tactic in inducing emotional states in cases when: (a) the subject has already invested significant effort to make a progress towards the final state of the given task, and has no possibility to easily instruct an undo command, and (b) the same incorrect operation was frequently repeated.

- Deliberate misunderstanding of subject's questions and providing useless, provoking answers (e.g., "you still haven't solved the task", "I am doing only what you are saying", "it is your task to solve", etc.).

- Pretending not to understand subject's utterance and asking for a repetition.

- Making comments aimed to increase the level of the stress in the subject (e.g., "it seems that you have made it more complicated", "the time is running out", etc.). These comments were especially effective when the subject was confronted with unsolvable tasks (cf. Figures B.8 and B.10 in Appendix B). The entire responsibility for the fail to solve these tasks was placed upon the subject.

In order to induce positive emotional states in the subject, the wizards simulated user-friendly behavior of the system, including:

- Recognizing verbose subject's utterances and performing requested operations even if they fall outside of the application's domain (e.g. allowing the subject to introduce and use a reference name for a graphical piece, etc.).
- Providing helpful comment and answers to subject's questions (e.g., proposing the next correct move, etc.).
- Making encouraging comments (e.g., "you are good", "excellent", "you have the best time", etc.).
- Capturing subject's image and displaying it as a part of graphical puzzles (cf. the Grid puzzle in Appendix B).

In addition, one of the tasks of the wizards was also to keep the verbal interaction between the subjects and the simulated system ongoing, in order that the subject can express emotions verbally. If the subject fell silent or refused to talk, the wizards tried to stimulate her to get back in the conversation (e.g., by reminding the subject that she can ask a question, etc.).

Apart from the "emotion stimuli" introduced above, there was not a predefined scenario that would strictly determine the behavior of the wizards (e.g., the time and the frequency of occurrences of simulated malfunctions, etc.). It was flexibly and dynamically adapted according to the actual situation in the session and the experience gained in previous experimental sessions.

After completing the test, subjects answered a questionnaire still not knowing that they participated in a WOZ simulation. In a post-session, the overall experimental settings were uncovered in order to debrief the subjects. Then subjects also participated in an informal interview. In the whole experiment almost 15 hours of session time is recorded[4].

2.6 Evaluation of the NIMITEK Corpus

Discussing the notion of ecological validity, Brewer (2000, p. 12) mentions that findings obtained with atypical populations (e.g., college students) in atypical settings (e.g., the laboratory) never have ecological validity until they are demonstrated to occur naturally in more representative circumstances. Therefore, it was necessary to evaluate the ecological validity of

[4] Please note that the NIMITEK corpus is available from the authors for research purposes upon request.

the NIMITEK corpus with respect to the guidelines introduced in Section 2.4. The evaluation of the emotional content of the NIMITEK corpus was performed in two phases. The primary aim of the first phase was to assess the level of ecological validity of the corpus. In addition, results of this phase served as point of departure for the second phase whose aim was to define a data-driven model of user states for the given WOZ scenario. In this chapter, we describe the first phase and discuss its results, because it directly relates to the question of ecological validity of the NIMITEK corpus. The second phase was performed for the purpose of implementing the adaptive dialogue management module in the NIMITEK prototype system. This phase is discussed in Chapter 4 (Subsection 4.4.4).

Three types of evaluators participated in the first phase. The first group (three German native speakers) was allowed only to hear audio recordings. These evaluators were influenced by lexical meaning as well. The second group consisted of three non-German speakers: two Serbian native speakers and one Hungarian native speaker (however, the last evaluator was born and living in Serbia, attending schools in Serbian language, etc.). These evaluators did not have knowledge of German language, have never lived in a German speaking environment, and did not have any contact with German language in everyday life. This group was also allowed only to hear audio recordings, but for this group the lexical meaning was missing and thus the prosody became central for evaluating emotions. A similar approach is used by Wendt and Scheich (2002) to evaluate the "Magdeburger Prosodie-Korpus" with respect to its emotional content. This corpus consists of two parts: linguistically meaningful German nouns and linguistically meaningless pseudo-words produced with different affective prosodies by an actor and an actress. The affective prosodies of the words contained in the corpus were assessed by a group of evaluators. The important difference is that they evaluated only short words of acted speech, while we consider genuine emotional expressions extended in time. Finally, one additional German native speaker was allowed to simultaneously hear and see video recordings from the NIMITEK corpus. All evaluators were naïve, i.e., without educational background that relates to the evaluation process (e.g., psychology, linguistics, sociolinguistics, etc.).

Four randomly selected sessions were evaluated in complete duration (approximately five hours), in order that evaluators take the history of interaction into account. An evaluation unit was a dialogue turn or a group of several successive dialogue turns. Only subjects' expressions were evaluated, while wizard's expressions were ignored. The total number of evaluated units was 424.

Evaluators performed this perception test independently of each other. They were given a starting set of "basic" labels {*joy*, *sadness*, *anger*, *fear*, *disgust*, *neutral*}, but they were also allowed to extend this set with additional labels, if necessary, according to their own perception. Thus, the choice of additional labels was data-driven. Preliminary inspection of the corpus showed that induced users' states are often graded and mixed phenomena. Therefore, evaluators were allowed to assign one or more labels to each evaluation unit. Recordings evaluated as emotional were further graded with respect to their intensity (three different levels: low, medium, high). After completing this phase of evaluation, the labels introduced by evaluators are classified in three groups:

- emotion labels,
- subject's state labels,
- talk style labels.

Since we are concentrated on examining certain aspects of affected speech rather then of facial gestures, we discuss evaluation results for the first two groups of evaluators (i.e., German native speakers and non-German speakers that were allowed only to hear audio recordings). We used majority voting in order to attribute labels to evaluation units. Table 2.1 shows all the introduced labels and the numbers of cases with majority voting for the first two groups of evaluators. We consider two kinds of majority voting:

- weak majority—exact two evaluators in a group agreed,
- strong majority—all three evaluators in a group agreed.

A total number of cases with majority voting is the sum of numbers of cases with weak and strong majority voting. Labels used by evaluators but with no majority voting (i.e., *fear* and *disgust*) are also included in the table. Absolute numbers of cases when these labels were assigned are given in Table 2.2.

Discussing the results of the first evaluation phase, we make several points.

(1) *Subjects signaled emotions overtly.* Confrontation to the simulated test proved to be a strong motivational factor. The combination of a motivating environment with already mentioned additional stimuli for an emotional response (e.g., intentional misunderstanding of subject's request, etc.) induced subjects to signal their emotions overtly. The evaluation of the

Table 2.1: Introduced labels and majority voting of German native speakers and non-German speakers that were allowed only to hear audio recordings. The number of evaluated units was 424.

Labels	German speakers majority voting			non-German speakers majority voting		
	total	weak	strong	total	weak	strong
Emotion						
anger	77	46	31	18	12	6
nervousness	8	8	-	**224**	131	93
sadness	8	7	1	1	1	-
joy	17	14	3	1	1	-
contentment	12	12	-	4	4	-
boredom	9	5	4	13	10	3
fear	-	-	-	-	-	-
disgust	-	-	-	-	-	-
neutral	**205**	124	81	54	45	9
Subject's state						
interested	2	1	1	1	1	-
surprised	6	4	2	22	16	6
insecure	26	19	7	71	60	11
disappointed	17	14	3	12	12	-
impatient	35	32	3	-	-	-
confused	30	21	9	-	-	-
accepting	8	5	3	-	-	-
pleased	2	2	-	-	-	-
stressed	47	43	4	-	-	-
thinking	10	10	-	-	-	-
Talk style						
commanding	53	47	6	137	94	43
off-talk	59	38	21	94	40	54
pedagogical	15	10	5	73	60	13
ironic	4	4	-	36	34	2

NIMITEK corpus shows that subjects signaled both positive and negative emotions. However, induction of negative emotions and emotion-related states was significantly more effective than induction of positive emotions and emotion-related states. The group of *positive* labels contains *joy*, *contentment* and *pleased*, as well as *interested* and *thinking* that can be considered positive with respect to subject's engagement to solve the given task. The other labels, except *neutral*, belong to the group of *negative* labels. According to majority voting results, German speaking evaluators attributed 10.14% of evaluation units with a positive label and 63.92% with a negative label, while non-German speaking evaluators attributed 1.42% of evaluation units with a positive label and 85.14% with a negative label.

(2) *Diversity of signaled emotions and their intensities.* Another property of the corpus is that it is not oriented to extreme representations of a few emotions only (for example: *anger*, *joy*, *fear*, etc.), but comprises also expressions of less intense, not *full-blown*, emotions (for example: *nervousness*, *pleased*, *insecure*, etc.). A convincing illustration of this fact is that non-German evaluators attributed 52.83% of evaluation units with *nervousness*, and only 4.25% with *anger*. As mentioned above, units attributed with an emotion label are further graded with respect to the intensity of the signaled emotion: low, medium or high. Table 2.2 shows the numbers of assigned emotion labels classified by intensity. In this table, we do not resort to majority voting, but give the absolute numbers of assigned labels.

Table 2.2: Numbers of assigned emotion labels classified by intensity of expressed emotion.

Emotion	***Low***	***Medium***	***High***
anger	248	144	30
nervousness	466	270	33
sadness	26	27	1
joy	66	41	3
contentment	130	23	3
boredom	142	27	1
fear	8	11	-
disgust	6	-	-

(3) *Emotional expressions are extended in modality.* Our experiment deals with the expressions of emotions in two modalities at a time: vocal and facial expressions. Although vocal expressions are prioritized in our re-

search, the laboratory settings give an opportunity to observe the correlation between these two modalities.

(4) *Emotional expressions are extended in time.* Producing recordings of emotional expressions that are extended in time is also an important requirement, because it allows different phases in development of emotions in subjects to be observed.

As mentioned above, an evaluation unit in this evaluation phase was selected as a dialogue turn or a group of several successive dialogue turns. It may contain utterances produced both by subjects and the wizard, as well as pauses in spoken dialogue while the system was performing instructed commands (we recall that only subjects expressions were evaluated, while wizards expressions were ignored). Therefore, the 424 selected evaluation units are rather long in duration—the average length of an unit is approximately 40 seconds. Such selection of evaluation units was necessary to demonstrate that emotional expressions are extended in time[5]. Evaluators were allowed to attribute one or more labels to each evaluation unit. Assigning more than one label to an evaluation unit means that there was a change in the expressed emotion within the observed evaluation unit (e.g., the occurrence of two different emotions or emotion-related states, or a change in the intensity of expressed emotion, etc.). According to the results of the majority voting, non-German evaluators attributed 52.12% (i.e., 221) of evaluation units with more than one label, while German evaluators attributed 39.86% (i.e., 169) of evaluation units with more than one label.

In the second evaluation phase—conducted for the purpose of implementing the adaptive dialogue management module—we used a finer selection of units. The same evaluation material was divided in 2720 evaluation units (see Chapter 4, Subsection 4.4.4).

(5) *Additional shared non-linguistic context.* In our experimental settings, the desktop of the subjects' PC was also recorded. It represented an additional non-linguistic context shared between the subjects and the simulated system. The subjects considered it to be a reliable source of information. In such cases when wizard's statements and actions were not in accordance with the actual state on the desktop, the subjects often considered the desktop to be more relevant. As a small illustration we give two examples. (a) After the wizard's instruction to solve a certain graphical puzzle, the subjects often immediately answered that the puzzle is already

[5] Moreover, as a general principle in language, larger units are to be considered because they function more directly in the realization of higher-level language patterns (Halliday, 1994, p. 19).

solved. In fact, the puzzle only appeared to be solved, and the second look at the desktop (cf. Figure B.10 in Appendix B) was enough that the subjects become aware of it. (b) After the wizard intentionally made a number of incorrect operations moving a graphical piece to left instead of to right, and vice versa, at least one subject concluded that she and the system have different perspectives on the desktop, and tried to address this problem by using phrases as ”my left” and ”my right”. In general, subjects may believe in shared linguistic context between them and the system. As Guindon (1988) notes, this could influence the language used by subjects. The inspection of all 6798 commands spontaneously produced by the subjects shows that non-linguistic context influenced the language of subjects to a high extent with respect to frequency of ”irregular” (e.g., elliptical or minor, etc.) utterances. We discuss this point in more detail in Chapter 3.

(6) *Different classes of non-neutral talking style.* Four classes of non-neutral talking style are marked in the obtained data: *commanding*, *off-talk*, *pedagogical*, *ironic.* Although they all carry information about speaker state and intention, one of them deserves a brief explanation. The *pedagogical* (could also be termed *teacherese*) talk refers to the kind of a language used by an examiner or an instructor trying to bring a subordinated listener to a certain conclusion or behavior by uttering a sequence of questions and instructions (therefore we decided to use the term *pedagogical*). The dialogue fragment between the subject and the system, given in Figure 2.3, illustrates this phenomenon. Solving the *Tower of Hanoi* puzzle, the subject tries to instruct the system to move the smallest disk from the first peg to the third peg. The dialogue is translated into English.

Considering all these points together leads to the conclusion that the first evaluation phase demonstrated a satisfying level of ecological validity of the NIMITEK corpus.

2.7 Conclusion

This chapter presented the NIMITEK corpus of affected behavior in speech-based human-machine interaction. It contains 15 hours of audio and video recordings produced during a refined Wizard-of-Oz experiment designed to induce emotional reactions. Ten native German speakers participated in the experiment. The language used in the experiment was German.

During the process of collecting the corpus proper attention was devoted to the issue of its ecological validity. We proposed two additional requirements that are to be met in order that a WOZ scenario designed to elicit

Subject: *The smallest ring from the one to the three ... Stop! Back ... Where should the smallest ring go? ... On the three, ... three! ... What are you doing there?*

Wizard: *I am doing just what you are asking.*

Subject: *Incorrect! ... Down ... these rings stay down! ... Understood?*

Wizard: *What rings?*

Subject: *The middle ring and the large ring, ... now put ...*

Wizard: *I don't understand you.*

Subject: *The smallest ring! ... On the three please, ... on the three ... No! On the three ... Where is the three? ... On the three! ... Where should the smallest ring be placed?*

Wizard: *On the three.*

Subject: *Please, do it!*

Wizard: *What should I do?*

Subject: *Pack the smallest ring on the three.*

Wizard: *I don't understand you.*

Subject: *The smallest ring ... on the three, ... on the three! ... No! Back! ... The smallest ring, ... on three, ... on three! ... Where is the three? ... On the three, ... three, ... three! ... On the three, ... correct!*

Wizard: *There is the three.*

Subject: *There it should go ... What are you doing? ... What are you doing?*

Wizard: *I am doing just what you are asking.*

Subject: *That is a lie!*

Wizard: *Please define.*

Subject: *You are not doing what I ask!*

Wizard: *I am doing just what you are asking.*

Subject: *Not true at all!*

Figure 2.3: Dialogue fragment illustrating *pedagogical* talk.

affected behavior could result in ecologically valid data. First, subjects have to be motivated to accomplish a given task in order that a successful accomplishment or a failure to accomplish could induce an emotional state. This requirement is introduced to address the problem of non-engaged subjects that are role-playing only. The second requirement for the successful emotion elicitation in WOZ experiments is that subjects have to be stimulated to express their emotions overtly. In addition, we discussed a need for a more sophisticated approach to dialogue management, concluding that experimental settings should allow experimenters to observe and control the development of the dialogue between subjects and the simulated system. Implications of this observation on wizard's dialogue strategies and response time were considered. Also, possible pitfalls of the proposed requirements were discussed.

Further, we described the experimental settings of the WOZ simulation conducted in the framework of the NIMITEK project and the corpus of affected behavior obtained in the experiment. Confronting subjects to the combined test of both intelligence and communication ability proved to be a strong motivating factor. The combination of a motivating environment with additional stimuli for an emotional response (e.g., intentional misunderstanding of subject's request, provocative interventions, etc.) induced subjects to signal their emotions overtly. Allowing them to address the system only with spoken instructions and questions, and advising that formulation of their utterances is also a part of a test caused that subjects used different strategies (e.g., changing prosody, producing elliptical utterances, etc.) to signal their emotions overtly.

The evaluation of the NIMITEK corpus with respect to its emotional content demonstrated a satisfying level of ecological validity. We summarize evaluation results in the following points. The corpus contains recordings of genuine emotions that were overtly signaled. It is not oriented to extreme representations of a few emotions only but comprises also expressions of less intense, everyday emotions. Emotional expressions of diverse emotions are extended in modality (voice and facial expression) and time. In addition, different classes of non-neutral talking style are marked in the obtained data.

The NIMITEK corpus had an important role in developing the dialogue management module in the NIMITEK prototype system. Here, we briefly indicate two main lines of research represented in this thesis that were supported by the NIMITEK corpus:

(1) *Modeling attentional information.* As mentioned above, in our experimental settings the desktop of the subjects' PC represented a non-linguistic context shared between the subjects and simulated system, and, thus, influ-

enced subjects' language. The analysis of subjects' utterances described in Chapter 3 showed that subjects often produced "irregular" (e.g., elliptical or minor, etc.) utterances. Thus, there was a need to develop structures and algorithms that support system's decision making processes when it is confronted with such user inputs. This is discussed in more details in Chapter 3. It introduces the concept of *the focus tree* in order to model attentional information on the level of a user's command and the rules for transition of the focus of attention for different types of user's commands.

(2) *Introducing an adaptive dialogue strategy for supporting users.* Resorting to the NIMITEK corpus, in Chapter 4 we introduce an adaptive dialogue strategy for supporting users while they solve a graphical task. It is aimed to address the negative user state on the two tracks: (i) to help a frustrated user to overcome the problem occurred in the interaction (e.g., problems related to the task itself or to the interface language, etc.), and (ii) to motivate a discouraged or apathetic user. The central idea is that the dialogue strategy is dynamically adapted according to the current state of the interaction.

These lines of research were integrated in the conceptual design and implementation of the dialogue management module in the NIMITEK prototype system.

Chapter 3

Modeling Attentional Information

3.1 Introduction

One of the widely accepted postulates of human-machine interaction is that it should be as natural as possible. An important aspect of naturalness of the interaction is certainly naturalness of the interface language. The essence of naturalness of the interface language is that users can express themselves without conscious effort to follow rules of a predefined grammar while producing their utterances. Forcing users to always produce "regular" utterances would be too restrictive and not well accepted. It can not be expected that users—and especially users in affected states—will always behave cooperatively and produce utterances that fall within the application's domain, scope and grammar. This implies that a language interface should be able to cope with various dialogue phenomena related to the users' language, such as different syntactic forms of users' utterances (from syntactically very simple utterances to verbose utterances), high frequency of ungrammaticalities, use of ellipses and anaphora, context dependent utterances, etc.

The aim of this chapter is to propose an approach to processing of "irregular" user's utterances. More precisely, this chapter introduces and discusses an approach to processing of the user's commands in human-machine interaction for the restricted model of commands contained in the NIMITEK corpus. However, as it is clear that this approach is not intended to cover a general case of unrestricted dialogue, it should also not be understood that this approach is limited to commands from the NIMITEK corpus only.

It covers the class of spoken dialogue systems that are intended to control graphical user interfaces, e.g., manipulating with graphical entities represented on the display, controlling graphical menus, solving graphically-based tasks and playing interactive board games that includes spatial reasoning, etc.

We make here an important note. This chapter primarily considers the question how to understand the user's utterances, e.g., to understand which move was instructed by the user while she solves a graphical puzzle. It does not consider implications that understanding of the user's utterances may have on the dialogue situation, e.g., it does not consider questions how does the user's move change the state of the puzzle, whether it is useful or legitimate, does it draw back the state of the puzzle from the expected final state, etc. These questions are considered in the next chapter.

Let us now recall that in the settings of the WOZ experiment the set of instructions accepted by the simulated system was not predefined. Detection and formulation of instructions were imputed to be a part of the test as an additional stimulus for subjects to express themselves verbally. Thus, the subjects had the freedom to formulate their own instructions spontaneously. We mentioned also that the desktop of the subjects' PC represented a non-linguistic context shared between subjects and simulated system. We gave two examples showing that subjects consider it to be a reliable source of information and indicated that it influenced their language. The inspection of the NIMITEK corpus shows that the subjects often produced elliptical or minor utterances. To illustrate this let us observe a dialogue fragment from this corpus shown in Figure 3.1. It includes a sequence of commands produced by the subject and performed by the system. The system did not produce verbal output. In the given dialogue fragment the subject solves the Tangram puzzle. The goal of this puzzle is to form a specific shape by using seven Tangram 2-D objects (e.g., triangles, etc.). Two kind of action over pieces were allowed: translation and rotation in the plane. The screenshot of the desktop representing the starting state of the puzzle is given in Figure 3.2.

> Big triangle ... rightwards ... rightwards ... stop ... downward ... stop ... rotate ... stop ... rightwards ... stop ... downward ... stop ... to the left ... stop ... thanks ...

Figure 3.1: Solving the Tangram puzzle: A sequence of commands produced by the subject.

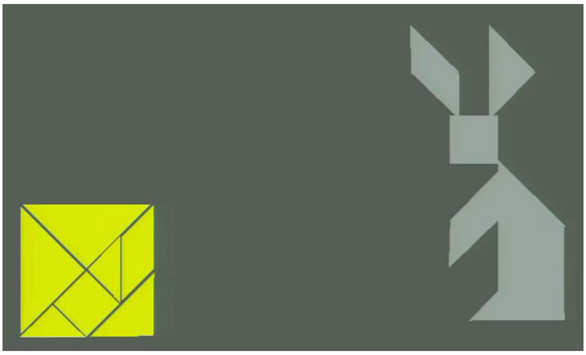


Figure 3.2: The screenshot of the desktop of the subject's PC. Tangram pieces are on the left side of the desktop, the shape to be formed is on the right side.

From the subject's point of view, the interaction could be summarized as follows. In the first command, she selects a Tangram piece. Afterwards, the subject assumes that the selected piece is a part of the shared knowledge between her and the system. Thus, until the end of the given fragment, she instructs only actions that should be performed over the selected piece, without explicitly referring to the selected piece itself. Consequently, utterances produced by the subject are elliptical—she omits to utter information that is already known by the system and, in the same time, brings new information in the focus of attention.

From the system's point of view, elliptical forms of commands have the advantage that the number of functional elements that the system has to process is reduced. The disadvantage is that such commands may be ambiguous. In the first command, the subject selects a big triangle piece. However, there are two same triangle pieces that could be identified as big. At this point there is no difference what piece of those two is selected. But, it does not hold in general—if one of those two pieces had been already positioned on the pattern shape, the distinction between them would have been important. The second command, "rightwards", is also ambiguous. The subject does not specify whether the selected triangle should be translated or rotated, although the command could be related to both of these actions. Again, the system is expected to conclude what operation is to be performed. Such commands are examples of "irregular" forms of utterances that occur in the NIMITEK corpus. We discuss this in more detail in Section 3.3. Thus, there is a need to develop structures and algorithms that support system's decision making processes when it is confronted with such user inputs.

Attentional information is already recognized as crucial for processing of utterances in discourse (Grosz and Sidner, 1986). This is discussed in Sections 3.4 and 3.5. In these sections, we investigate how attentional information can be used to process subjects's commands of different syntactic forms: We model attentional information on the level of the user's command and introduce rules for transition of the focus of attention.

3.2 Background and Related Work

State-of-the-art researches in dialogue management and underlying theoretical concepts have been already sufficiently covered in various reports (cf. Wilks et al. 2006, Catizone et al. 2002, Xu et al. 2002). Approaches to dialogue modeling are usually classified in two groups: dialogue grammars (i.e., pattern-based) and plan-based. The former approach is based on attempts to identify and represent surface patterns of dialogue (e.g., adjacency pairs, etc). The latter approach attempts to explicitly represent the goal of the task in a given interaction. Its point of departure is that the speaker's speech act is a part of a plan and that it is the listener's job to identify and respond appropriately to this plan.

This section provides a particular view in some theoretical insights from Conversational Analysis and some of their implementations in well-known existing spoken dialogue systems. We use it to motivate our approach to dialogue management that is somewhat different from these dominant approaches.

An observation that underlies the use of dialogue grammars to parse the structure of a dialogue is that there are sequencing interrelations between dialogue acts which are called *adjacency pairs* (Schegloff, 1968). An adjacency pair is a unit of conversation that contains two sequent dialogue turns produced by two speakers, where the second dialogue act is an appropriate response to the first. Typical examples of such two-part structures are question followed by answer, greeting followed by greeting, speech acts that require acceptance or rejection on the part of a hearer (e.g., offers, proposals, bets, invitations) followed by rejection or acceptance, etc.

Roulet (1992) goes a step further arguing that relations within a dialogue instance do not concern single dialogue acts, but more complex entities, comprising several dialogue acts, that he calls *moves*. His reasoning is closely connected to the notion of intentionality. In numerous theoretical and practical approaches to the question of the nature of dialogue, the

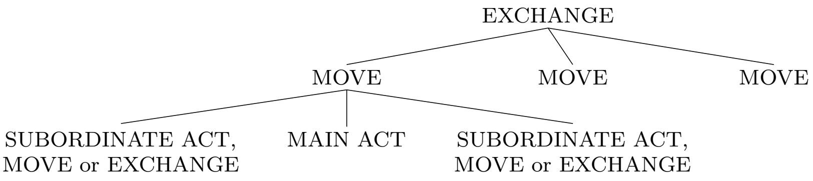

EXCHANGE
MOVE
MOVE
MOVE
SUBORDINATE ACT, MOVE or EXCHANGE
MAIN ACT
SUBORDINATE ACT, MOVE or EXCHANGE

Figure 3.3: Hierarchical structure of the linguistic exchange (adopted and adjusted from the original work of Roulet 1992, p. 97).

notion of intentionality is one of the most essential points. Whereas individual intentionality of a speaker uttering an isolated speech act is sufficiently elaborated, this notion becomes more complex at the level of a conversation that involves more participants. The observation that appears to be widely accepted is that intentionality is not given at the beginning of a conversation—it evolves as the conversation proceeds (Grosz and Sidner 1986, Searle 1992a, Roulet 1992). Thus, for Roulet (1992, p. 94), the intentionality shared between the participants is a constant object of *negotiation* between them. Since this activity of negotiation can be perceived in all the phases of a conversation, he makes a hypothesis that the activity of negotiation determines the structure of a verbal exchange. Roulet states that any negotiation consists of at least three phases: a *proposition*, a *reaction* and an *evaluation* phase. In addition, he recognizes the concept of recursivity in the development of a negotiation. One source of recursivity relates to the completeness of a negotiation phase (*interactive complétude constraint*). A participant in a conversation may provide an appropriate reaction only if the proposition is clear and complete. If not, the listener in her turn may open a secondary negotiation, in order to get the additional information that she needs. Upon successful closure of a secondary negotiation, both participants can get back to the main negotiation. Roulet illustrates this with an example: If an itinerant dealer offers him a carpet, he can react by an acceptance or a rejection only if he knows a price. If the dealer does not mention the price in the offer, Roulet will have to open another, secondary, negotiation to get the price. When this secondary negotiation is closed (i.e., when the dealer gives the price), they can get back to the main negotiation (i.e., Roulet could accept or reject to buy a carpet). Considering the structure of conversation as negotiation, Roulet (1992, p. 97) suggests that the *linguistic exchange* has a hierarchical structure, with constituents at the two levels, as shown in Figure 3.3 that is adopted and adjusted from his original work.

This hierarchical model of the linguistic exchange influenced existing models of dialogue structure in human-machine interaction. Bilange (2000) reports the dialogue management module in the SUNDIAL system, a spoken dialogue system designed to handle travel conversations (e.g., flight reservations, train table information, etc.) Analyzing dialogue corpora on flight reservations, Bilange (2000, p. 190–1) defines a structural description of dialogue that consists of four levels: the transaction level (i.e., a sequence of exchanges), the exchange level (i.e., a *negotiation* about a topic), the intervention level (i.e., interventions that carry an illocutionary function of *initiative*, *reaction* or *evaluation*), and the dialogue act level. Similarly, in the speech translation system Verbmobil, Alexandersson and Reithinger (1997, p. 2231–2) divide the intentional structure[1] into four levels: the dialogue level, the phase level (distinguishing three dialogues phases: greeting, *negotiation*, closing), the turn level (the main turn classes for negotiation dialogues are: *initiative*, *response*, *transfer-initiative*, and *confirmation*), and the dialogue act level. Common for both these models of the dialogue structure, besides that they aim to describe interactive nature of dialogue exchanges, is task-orientation. Introducing the model used in the SUNDIAL system, Bilange says:

> It can characterize the units that compose a dialogue. This characterization is twofold: it identifies a dialogue entity of a particular type and playing a certain role (Bilange, 2000, p. 189).

.

Introducing the plan processor—i.e., the component of the Verbmobil system responsible for the construction of the intentional structure and for the recognition of dialogue constituents—Alexandersson and Reithinger note:

> The plan hierarchy is compiled off-line into a context-free grammar [...] it is used not to plan actively, but to recognize plans (Alexandersson and Reithinger, 1997, p. 2231).

The idea that underlies both the dialogue models can be summarized in the two following points:

[1] Alexandersson and Reithinger (1997, p. 2231) use the term *intentional structure* to denote *a tree-like structure mirroring different abstraction levels of the dialogue (cf. dialogue phase, turn).* This is certainly not an unappropriate term, but it may be seen as too general for the context in which it is used. For the purpose of clarity, we point out that the same structure may be also denoted using a more specific term—*task structure*—that represents a special case of the intentional structure (Grosz and Sidner, 1986, p. 180).

- Task structure of the dialogue is predefined and a task-specific role is assigned to each dialogue constituent (e.g., initiative, response, confirmation, etc.).

- At run time, the dialogue manager takes an ensuing user's speech act and tries to assign one of predefined roles to it. In other words, the dialogue manager tries—according to a predefined plan of the interaction development—to map user's speech acts onto a set of predefined dialogue constituents.

Such a determination of an ensuing speech act in terms of how well it matches a predefined task-specific role implies the task-orientation of these models. Our approach is somewhat different from these popular approaches to dialogue management that are primarily concentrated on the task structure. In contrast to dialogue models based on the task structure, we concentrate on the attentional information in the dialogue. Thus, this overview was just a point of departure for further discussion on attentional state in human-machine interaction that is given below in the chapter. In the next chapter, we discuss how our model of attentional state and the state of the task can be integrated as parts of the state of the interaction.

3.3 Annotation of Dialogue Acts

In order to examine various forms of utterances produced by the subjects, we annotated dialogue acts in the NIMITEK corpus. Dialogue acts are basic elements of human interaction in the sense that interaction can be generally perceived as a sequence of exchange dialogue acts. However, interaction is not merely a sequence of dialogue acts. One of the most obvious principle that can be found in interaction is that each dialogue act creates a space of possibilities of appropriate response dialogue acts (Searle, 1992a, p. 8). Interrelations between dialogue acts are certainly not trivial. Moreover, the research question of getting an account that gives constitutive rules for conversation in general is still not answered. Therefore, most of dialogue act schemes used in existing spoken dialogue systems nowadays are task-oriented (Alexandersson et al., 2000, p. 442) and thoroughly elaborated (e.g., the dialogue act scheme used in the Verbmobil project differentiates between more than 30 different dialogue acts, cf. Alexandersson et al. 1998, p. 19). Here, in contrast to this trend, we start with a more fundamental—and consequently less task-oriented—classification of dialogue acts introduced by Halliday (1994).

Table 3.1: Subjects' utterances illustrating speech functions.

Speech function	***Example***
Command	"Rotate to the left."
Offer	-
Question	"What are names of these rings?"
Statement	"You are not doing what I say."

Considering the nature of dialogue, Halliday (1994, p. 68–71) suggests an interpretation of the clause in its function as an *exchange*. He distinguishes between two fundamental types of speech role—giving and demanding—as well as between two basic types of the exchange commodity—verbal (*information*) and nonverbal (*goods-&-services*). The role in the exchange and the exchange commodity define the four primary speech functions of: *command* (demanding goods-&-services), *offer* (giving goods-&-services), *question* (demanding information) and *statement* (giving information). We adopt this classification of dialogue acts. Table 3.1 gives examples of subjects' utterances from the NIMITEK corpus that illustrate these speech functions. The entry in the table that corresponds to offers is empty. According to the experimental settings, the subjects were allowed only to verbally address the system. Thus, no offers produced by the subjects (e.g., body and facial gestures, etc.) were annotated. It does not mean that there were no offers produced by the subjects in the NIMITEK corpus—they are just not annotated because they are not in the focus of our attention in this thesis.

A question that comes up is whether such a simple classification based on speech roles is adequate for the purpose for which the research was undertaken in the first place—examining various forms of utterances produced by the subjects while they interacted with the system. The answer is positive and we briefly discuss it. Although Halliday (1994, p. 19) particularly concentrates on the clause, it is important to note that he does not observe just a clause *in abstracto*, i.e., a clause isolated from the surrounding dialogue context. He points out that a correlation between the speech roles (i.e., giving and demanding) also supports the concept of dialogue context (Halliday, 1994, p. 68-9). Halliday notes that giving means inviting to receive, and demanding means inviting to give. In his words, when a speaker adopts for himself a particular speech role, he also assigns to the listener a complementary role. In the example given in Table 3.1, when the subject says "What are names of these rings?", he requires the system to take on

Table 3.2: Results of the annotation of verbal dialogue acts produced by the subjects in the NIMITEK corpus.

Speech function	***#Spontaneously produced***	***#Predefined***	***#All***
Command	6798 (74.93%)	161 (1.77%)	6959 (76.57%)
Question	390 (4.29%)	0 (0%)	390 (4.29%)
Statement	1727 (19.00%)	13 (0.14%)	1740 (19.14%)
Total	8915 (98.09%)	174 (1.91%)	9089 (100%)

the role of the information supplier. In its turn, moving into the role of speaker, the system has the opportunity to adopt the required role for itself. This complementarity of the introduced speech roles supports the interpretation of the clause as an interactive event. It is clear that the system can respond in different ways depending on its dialogue strategy. However, in its response it adopts for itself one of the two mentioned speech roles, even if it might not match the role required by the subject in the previous turn. In this sense, Halliday's classification of dialogue acts served as a point of departure—it gave us an useful overview and directed towards an additional annotation (cf. Subsection 3.3.1).

The annotation process was performed by a group of instructed student annotators. They used the described classification (i.e., command, question, statement) to annotate all verbal dialogue acts produced by the subjects in the NIMITEK corpus. The results of the annotation process are summarized in Table 3.2. The given numbers relate to utterances that were spontaneously produced as well as to utterances that were predefined (i.e., "Start the test", "Task completed" and "I give up. Next task."). However, in examining forms of subjects' utterances, we consider only those utterances that were spontaneously produced by the subjects.

We mention three important implications of these results. First, 98.09% of subjects' verbal dialogue acts were spontaneously produced. In previous chapter, we said that one of the requirements for a successful emotion elicitation in the performed WOZ experiment is that subjects have to be stimulated to express themselves. Such a high percentage of spontaneously produced utterances is an evidence of the fulfillment of this requirement. Second, 74.93% of subjects' verbal dialogue acts are spontaneously uttered *commands*. Therefore, we devote a particular attention to this class of utterances. Annotation of commands is described in the coming subsection. Finally, the third implication relates to the dialogue strategy applied by the wizard. According to the experimental settings, problems in the interaction were caused on purpose and the evaluation of emotional content demonstrated that subjects expressed negative emotions overtly. Still, questions make only 4.29% of all utterances produced by the subjects. In addition, the subjects demanded support from the system in only 12 of 6798 commands (cf. Table 3.3), although the human operator playing the role of the system offered support 59 times explicitly using the word *help*, e.g., *Do you need help?* (in German: *Brauchen Sie Hilfe?*). Thus, a dialogue strategy aimed to support the user to overcome problems that occur in the interaction must not rely on the assumption that the user will clearly state a need for support. The system should rather detect such a need and be initiator and carrier of provided support. This issue is further discussed in Chapter 4.

3.3.1 Annotation of Commands

The class of commands spontaneously uttered by the subjects is the most represented class (74.93%) of subjects' verbal dialogue acts in the NIMITEK corpus. Therefore, we perform additional, more detailed annotation of the subjects' commands. In contrast to Halliday's classification that is of a general nature, here we use a corpus-specific classification. We create this classification based on the inspection of the NIMITEK corpus and observations on the structure of spoken language made by Campbell (2006).

Considering the structure of spoken language, Campbell differentiates between two types of content that can be signalled in an utterance:

> [...]we distinguished each utterance as being either of I-type or A-type content; the former primarily expressing propositional content (or *Information*[])], and the latter primarily expressing *Affect*. (Campbell, 2006)

Both types are often simultaneously signalled in spontaneous spoken language. Campbell introduces the notions of *fillers* and *wrappers* to denote parts of utterances that relate to these two types of content. The term *filler* is used to describe the information content of an utterance, while the term *wrapper* is used to describe the affect portions of an utterance. In Campbell's words:

> Whereas in written communication the word sequences are usually carefully deliberated and well-formed, in the case of spontaneous speech the flow is generated in real-time and a stream of

> words and phrases might typically (in colloquial English) appear as follows:
>
> " ... *erm, anyway, you know what I mean, ..., it's like, er, sort of* **a stream of** *... er ...* **words, and phrases**, *all* **strung together**, *if you know what I mean, you know ...* "
>
> where the words in bold-font form the content (or the *filling* of the utterance) and the italicised words form the *wrapping* or decoration around the content. (Campbell, 2006)

Keeping these observations in mind, we conducted an inspection of commands from the NIMITEK corpus. As expected, they often contain words or phrases that explicitly relate to entities from the currently salient focus space. For example, in the case of the Tower of Hanoi puzzle, those entities are rings and pegs. Some typical examples of such commands are:

- The first to the three. (Die Erste auf die Drei.)
- The two on three. (Die Zwei auf die Drei.)
- Rightwards. (Nach rechts.)
- The one rightwards. (Die Eins nach rechts.)
- The next ring. (Den nächsten Ring.)

In general, a fully formulated command in the Tower of Hanoi puzzle is expected to contain following information: which disk should be moved, and to which peg it should be moved. As illustrated, a command may contain only a part of this information, e.g., information only about the disk or only about the peg. Commands "Rightwards" and "The next ring" are in this sense elliptical. However, commands may also contain some additional information that does not directly relate to propositional content (e.g., phrases of courtesy, etc.). Some such examples taken from the NIMITEK corpus are:

- *I would like to put* **the smallest disk** *on* **the three**. (*Ich würde gern'* **die kleinste Scheibe** *auf* **die Drei** *legen.*)
- *I would like now to move* **the disk three** *on* **peg three**, *which does not work.* (*Ich möchte jetzt gern'* **die Scheibe Drei** *auf* **Turm Drei** *verschieben, was nicht funktioniert.*)

- **The middle disk** *please on* **the number two**. (**Den mittleren Ring** *bitte auf* **die Nummer Zwei**.)
- *On* **the three** *please.* (*Auf* **die Drei**, *bitte.*)

Words and phrases that relate to propositional content (i.e., fillers) are given in bold. The words given in italic represent wrappers in the sense of Campbell (2006) —"decoration" around the propositional content.

From the system's point of view, fillers are important for understanding propositional content. In our approach, we refer to them using the term—*focus instances*[2]. The next sections consider in more detail their role in processing of utterances by the system. On the other hand, since wrappers carry affect information, they are important for recognition and tracking of the user's emotional state from linguistic information. This issue is illustrated in Appendix A. As we show in the rest of this subsection, commands containing focus instances (as those illustrated above, both with and without wrappers) represent an important group of subjects' commands in the NIMITEK corpus.

During the inspection of the NIMITEK corpus, we also recognized commands that do not contain focus instances. We classify them as follows:

- Undo commands—Commands from this class reverse effects of previously instructed command. Some examples of such commands are:
  - Please undo. (Mach' das bitte rückgängig.)
  - Back. (Zurück.)
  - Please go back. (Geh' bitte zurück.)
- Redo commands—These commands explicitly instruct a repetition of previously instructed command. For example:

[2] We choose the term *focus instance* to correspond with the tradition in linguistic theory of denoting information structure of discourse. Observing phonological constituency in language, Halliday (1994, p. 292) shows that prosodical patterns serve to organize discourse into information units, with each information unit comprising the functions of Given and New. The Given is information that is presented by the speaker as recoverable—something that is not news (e.g., something that has been mentioned before, something inherent in a given context, etc.). The New is information that is presented by the speaker as non-recoverable—something that is news (e.g., something that has not been mentioned previously, something unexpected, etc.) (Halliday, 1994, p. 298). Each information unit is organized as a pitch contour and the New is marked by prominence (i.e., it carries the main pitch movement) (Halliday, 1994, p. 296). In Halliday's words, the New carries information FOCUS. Similarly, we use the term *focus instance* to denote dialogue act constituents (not necessarily marked by the tonic prominence) that carry new or salient information.

  - A bit further. (Ein Stück weiter.)
  - Please more. (Mehr bitte.)
  - Again. (Noch einmal.)

- Stop commands—Commands from this class explicitly instruct a termination of previously instructed command. Typical examples of such commands are:

  - Stop! (Stopp!)
  - No, stop! (Nein, halt!)
  - Don't rotate! (Nicht drehen!)

- Commands specifying a *focus class*—We illustrate this class of commands in the context of the Tangram puzzle. Five out of seven Tangram pieces are triangles (cf. Figure 3.2). In the command that contains a focus instance (given in bold):

  - I take **the big triangle on top**. (Ich nehme **das große Dreieck oben**.)

  the subjects exactly specifies which triangle should be selected. In contrast to this, an example of a command specifying a *focus class* is:

  - I take the triangle. (Ich nehme das Dreieck.)

  In this command, the subject just specifies that a triangle should be selected. Although the subject uses the definite article *the* (*das*) which signals that she refers to a particular triangle, she does not provide additional information that could help the system to conclude which triangle should be selected. We use the term *focus class* to represent a generalization of a subset of focus instances present in a given dialogue context.

- Commands containing ellipsis-substitutions— Ellipsis-substitution is a form of anaphoric cohesion in a discourse, *where we presuppose something by means of what is left out* (Halliday, 1994, p. 316). To illustrate this, let us observe the following sequence of questions:

  > Why are you *moving* it on peg 2? Why? Why are you *doing* this step? (Warum *fährst* du auf Säule 2? Warum? Warum *machst* du diesen Schritt?)

Table 3.3: Results of the annotation of spontaneously produced subjects' commands in the NIMITEK corpus.

Type of command	***#Occurences***
Commands containing focus instances	5469 (80.45%)
Undo commands	65 (0.96%)
Redo commands	300 (4.41%)
Stop commands	817 (12.02%)
Commands specifying a focus-class	125 (1.84%)
Commands containing ellipsis-substitutions	10 (0.15%)
Help commands	12 (0.18%)
Total	6798 (100%)

The last question contains as ellipsis-substitution—the verb *do*. The subject replaces the verb *move* (*fahren*) with the general verb *do* (*machen*). Typical examples of commands with ellipsis-substitutions are:

- Please do it! (Bitte tu' das!)
- Do what I say! (Tu', was ich sage!)

Isolated from the surrounding dialogue context, these utterances do not explicitly carry information what is the system expected to do.

- Help commands—This class includes commands in which the user explicitly asks for support, such as:
  - Help. (Hilfe.)
  - I would like help. (Ich hätt' gern' Hilfe.)
  - I said: help! (Ich sagte: Hilfe!)
  - I need help in the communication with the system. (Ich brauche Hilfe in der Kommunikation mit dem System.)

After the inspection of the NIMITEK corpus, we differentiated between seven classes of spontaneously uttered subjects' commands introduced above. This classification was used to annotate all commands from the NIMITEK corpus that were spontaneously produced by the subjects. The annotation process was performed again by instructed student annotators. The numbers of occurrences of commands from each class is given in Figure 3.3.

According to the annotation results, commands that contain focus instances are most dominant, i.e., 80.45% of all spontaneously produced commands belong to this class. Thus, this class has a central position in our approach to processing of the user's commands. The next sections introduce this approach. The other classes of commands (i.e., those that do not contain focus instances, e.g., undo, redo, stop, etc.) are then also taken into account.

3.4 Attentional Information

The theory of discourse structure introduced by Grosz and Sidner (1986) is closely related to two nonlinguistic notions: intention and attention. Whereas intentions occupy a central position in explaining discourse structure, attention is denoted as *an essential factor in explication the processing of utterances in discourse* (Grosz and Sidner, 1986, p. 175). Therefore, we consider the notion of attention in more detail. Grosz and Sidner note:

> Attentional state contains information about the objects, properties, relations, and discourse intentions that are most salient at any given point. It is an abstraction of the focus of attention of the discourse participants; it serves to summarize information from previous utterances crucial for processing subsequent ones, thus obviating the need for keeping a complete history of the discourse. (Grosz and Sidner, 1986, p. 177)

Grosz and Sidner model the attentional state by a set of *focus spaces.* They call the collection of focus spaces available at any one time the *focusing structure* and the process of manipulating spaces *focusing.* In the focusing process introduced by Grosz and Sidner, a focus space is assigned to each discourse segment. A focus space contains entities that are salient in the given discourse segment (e.g., entities that have been mentioned explicitly in the segment or introduced implicitly in the process of producing or comprehending the utterances in the segment, etc.) (Grosz and Sidner, 1986, p. 179). This is illustrated in Figure 3.4 that is adopted and adjusted from the original work of Grosz and Sidner (1986, p. 180–1).

The intentional structure of the discourse in the given example, including relationships among discourse segments purposes, is represented in the dominance hierarchy on the left in the figure. Discourse segment DS1 dominates discourse segments DS2 and DS3. The focusing structure is given on the right in the figure. Each of these discourse segments is tied to a focus

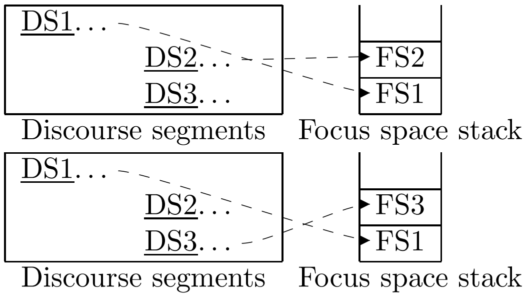


Figure 3.4: Discourse segments and focus spaces (adopted and adjusted from the original work of Grosz and Sidner 1986, p. 181).

space. The state of focusing when discourse segment DS2 is being processed is given in the first part of Figure 3.4. Being the most salient, focus space FS2 is positioned on the top of the stack. Focus space FS1, assigned to the dominating discourse segment DS1, is also accessible, although less salient. When discourse segment DS3 is being processed, focus space FS2 has been popped from the focus space stack, and focus space FS3 has been pushed onto it.

Grosz and Sidner (1986, p. 182-192) provide concrete, well-elaborated examples for illustration of their theory (e.g., for an argument from a rhetoric text, for a task-oriented dialogue, etc.). They note that their theory, although still incomplete, does provide a solid basis for investigating both the structure and meaning of discourse, as well as for constructing discourse-processing systems. They also suggest research problems of primary importance that remain to be further explored. One of them is: *Investigation of alternative models of attentional state.* (Grosz and Sidner, 1986, p. 202)

This chapter addresses this research problem for a very restricted case. We move away from the level of general discourse (including various forms of spoken and written discourse such as task-oriented dialogues, everyday conversations, monologues, narrations, textbooks, newspaper articles, etc.) and consider only attentional information on the level of the user's command. Restrictions in our approach, compared with the approach introduced by Grosz and Sidner, result from the following assumptions:

- All focus spaces that become salient during the processing of the discourse segments are known in advance. In the given example, the focus spaces are {FS1, FS2, FS3}.

- There is a structural relation between focus spaces that corresponds

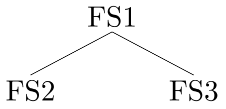


Figure 3.5: A simple focus tree.

to the dominance hierarchy of discourse segments. In the example, discourse segment DS1 dominates discourse segments DS2 and DS3. Therefore, we introduce a structural relation according to which focus space FS1 "dominates" focus spaces FS2 and FS3. We use a tree structure—the *focus tree*—to model attentional information. A simple focus tree that corresponds to the given example is shown in Figure 3.5. Whereas the focus space stack represents a collection of available focus spaces at given point, the focus tree is determined beforehand and fixed.

It is clear that these very restricting assumptions do not hold for the case of general discourse. Neither are focus spaces always known in advance nor can their structural relationship always be represented in a tree structure. Nevertheless, if we consider the restricted model of commands produced by the subjects in the NIMITEK corpus, these assumptions appears to be adequate, as we will show soon. But first we should explain the concept of the focus tree.

The focus tree preserves the idea of recursive development of the focusing structure introduced by Grosz and Sidner. They use a stack structure to represent the dynamic nature of the attentional state. The stacking and manipulating of focus spaces reflects the relative salience of the entities in each space (Grosz and Sidner, 1986, p. 180). In the focus tree, the dynamical nature of attentional state is represented by placing the *focus of attention* on one of nodes in the focus tree. This comparison is illustrated in Figure 3.6. It shows the sequence of states of the focus stack (Grosz and Sidner) and the focus of attention in the focus tree (introduced in our approach) during the processing of the discourse segments in the given example. Before the segments are processed (State 1), the focus stack is empty and the focus of attention is not placed on any of the nodes in the focus tree. When segment DS1 is being processed (State2), focus space FS1 is positioned on the stack. In the focus tree it is represented by placing the focus of attention on the node FS1 (the node is represented in oval). Processing of segment DS2 (State 3) pushes focus space FS2 on the top of the stack. Corresponding to

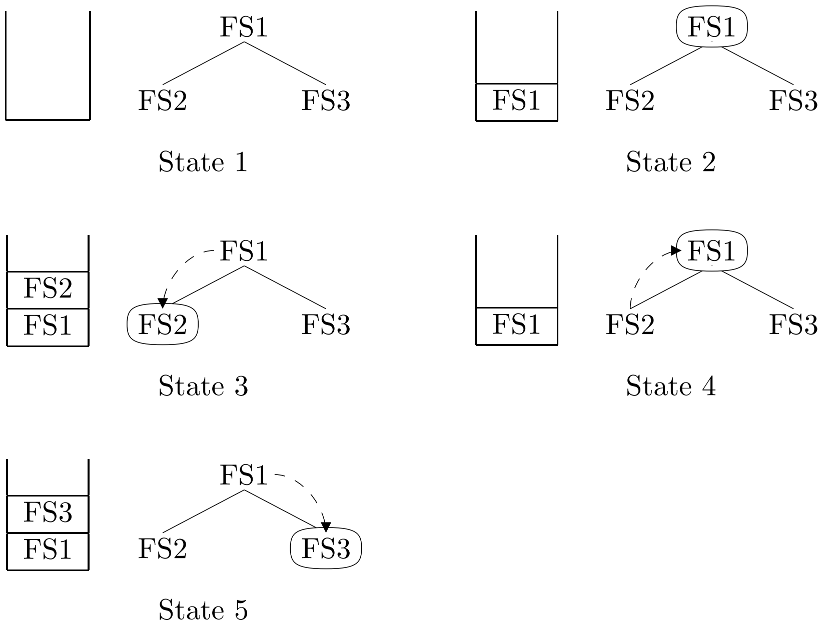


Figure 3.6: Comparison between focus stack and focus tree.

the fact that this focus space is the most salient at the moment, the focus of attention is shifted on the node FS2 in the focus tree. After segment DS2 has been processed (State 4), focus space FS2 has been popped from the stack. Focus space FS1 is now on the top of the stack and the focus of attention is again placed on the node FS1 in the focus tree. Processing of segment DS3 (State 5) gives rise to focus space FS3—pushing focus space FS3 on the stack is represented by placing the focus of attention on the node FS3 in the focus tree.

Starting from the introduced assumptions, it can be summarized that the focus tree encapsulates the set of all possible states of the focus space stack for a given discourse. States of the focus space stack are denoted by the position of the focus of attention in the focus tree: A node that carries the current focus of attention corresponds to a focus space that is placed on the top of the stack, its parent node corresponds to a focus space that is placed below, and so on—all ancestor nodes correspond to focus spaces contained in the stack, where the root node of the focus tree corresponds to a focus space placed on the bottom of the stack.

Now, when we briefly motivated the notion of the focus tree, we state an analogy that further explains the nature of the focus tree. Similarly as a dialogue instance contains dialogue acts (e.g., user's commands, etc.), a focus space—as it was defined by Grosz and Sidner (1986, p. 177)—contains entities that we refer to as focus instances (introduced in Section 3.3.1). Since in our approach to modeling attentional information we move away from the level of general discourse towards the level of the user's command, we actually deal with focus instances rather then with focus spaces. Therefore, we make a clarification: nodes in a focus tree do not correspond to focus spaces, but to focus instances. We illustrate this for a concrete example taken from the NIMITEK corpus—the Tangram puzzle. After inspection of subjects' commands from the corpus, we differentiate among four *focus classes* whose instances form attentional information. They are given in the following list, starting from the most general focus class and ending with the most specific:

- Task focus—Focus instances contained in this class relate to the tasks given to the subjects in the WOZ experiment described in Chapter 2 (e.g., the Tangram puzzle, the Tower of Hanoi puzzle, the Grid puzzle, etc.).

- Object focus—Focus instances contained in this class relate to graphical objects that can be manipulated in the given tasks (e.g., Tangram pieces, disks in the Tower of Hanoi puzzle, tiles in the Grid puzzle, etc).

- Action focus—Focus instances contained in this class relate to actions that can be performed over selected objects. For the Tangram puzzle, there are two focus instances contained in this focus class that relate to actions of translation and rotation, respectively.

- Direction focus—Focus instances contained in this class relate to further specification of actions that can be performed over selected objects. For the action of translation, there are 4 focus instances that relate to direction (up, down, left and right), and for the action of rotation there are two focus instances that relate to direction (clockwise and counterclockwise).

These focus classes are interrelated—an instance of a more specific focus class is a sub-focus of an instance of the immediately preceding more general focus class. We shortly explain a sub-focus relation: focus instance f_1 is a sub-focus of focus instance f_2 if focus instance f_1 cannot become salient in

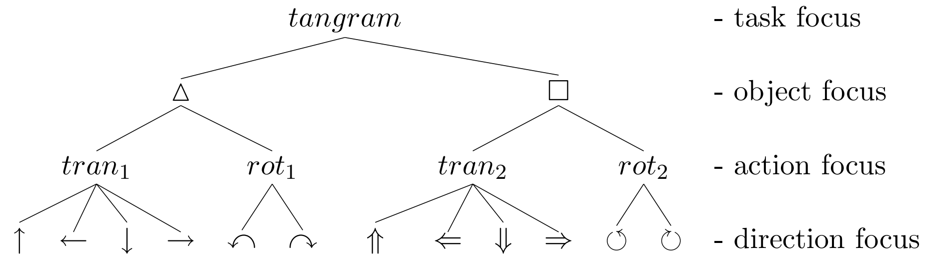


Figure 3.7: The simplified focus tree for the Tangram puzzle.

the given dialogue without f_2 being also salient in the same moment (f_2 may be explicitly mentioned in the dialogue or implicitly introduced into the dialogue context). For example, a focus instance representing an action over a Tangram piece is a sub-focus of a focus instance representing that Tangram piece because we have to specify the piece before we can perform an action over it. It is important to note that a sub-focus relation is a kind of semantic relation and not determined with the syntactical structure of users' utterances. Thanks to this property it is possible, as we discuss below, to utilize sub-focus relations to process the user's commands of different syntactic forms. Sub-focus relations are preserved in the focus tree for the Tangram puzzle. Each instance of these classes is represented by a node in the focus tree. Each node, except the root node, represents a sub-focus of its parent node. The root node represents the most general focus instance. Nodes at the same level of the tree belong to the same focus class. The focus tree for the Tangram puzzle is given in Figure 3.7. For the purpose of easier representation, we reduce the number of Tangram pieces to two: the triangle (△) and the square (□). It means that we show only a part of the "bigger" focus tree including all seven Tangram pieces that is important for the following discussion. However, this reduced representation implies by no means a reduction of complexity of the observed dialogue domain. Table 3.4 provides short descriptions of all focus instances in this focus tree.

At every moment of interaction, the current focus of attention is represented by exactly one node in the focus tree. Mapping of an ensuing user's command onto the focus tree is performed with respect to the position of the current focus of attention. Also, the user's command may change the focus of attention. This is considered in the next section in more detail.

Table 3.4: Focus instances in the simplified focus tree for the Tangram puzzle

Focus instance	***Focus class***	***Description of focus instance***
$tangram$	task	Tangram puzzle
$\triangle$	object	triangle
$tran_1$	action	translation of $\triangle$
$\uparrow$	direction	upward translation of $\triangle$
$\leftarrow$	direction	leftward translation of $\triangle$
$\downarrow$	direction	downward translation of $\triangle$
$\rightarrow$	direction	rightward translation of $\triangle$
rot_1	action	rotation of $\triangle$
$\curvearrowleft$	direction	counterclockwise rotation of $\triangle$
$\curvearrowright$	direction	clockwise rotation of $\triangle$
$\square$	object	square
$tran_2$	action	translation of $\square$
$\Uparrow$	direction	upward translation of $\square$
$\Leftarrow$	direction	leftward translation of $\square$
$\Downarrow$	direction	downward translation of $\square$
$\Rightarrow$	direction	rightward translation of $\square$
rot_2	action	rotation of $\square$
$\circlearrowleft$	direction	counterclockwise rotation of $\square$
$\circlearrowright$	direction	clockwise rotation of $\square$

3.5 Transition of the Focus of Attention

For easy reference, we introduce the following abbreviations:

- f — a focus instance,
- g — a node in the focus tree that represents f.

To give an example: the command "rotate to the left" includes two focus instances f_1 = "rotate" and f_2 = "to the left" that belong to the action and direction focus classes, respectively. Observed out of context, the focus instance f_1 can be represented by nodes $\{rot_1, rot_2\}$, while the focus instance f_2 can be represented by nodes $\{\leftarrow, \curvearrowleft, \Leftarrow, \circlearrowleft\}$ in the focus tree given in Figure 3.7. For a node g, we define rank $R(g)$ as the length of the path from the root node to the node g. In addition, let us assume that all focus instances (if there is any) contained in a command belong to different focus classes. This assumption does not hold in general. The command "move triangle and square rightwards" contains two focus instances "triangle" and "square" that belong to the object focus class. However, each such a command can be divided in a sequence of commands whose all focus instances belong to different focus classes, e.g., "move triangle rightwards" and "move square rightwards".

In the following subsections, we introduce and illustrate algorithms for transition of the focus of attention for the restricted model of actions represented as frames, as in Section 3.4. Rules for transition of the focus of attention are illustrated for commands from the NIMITEK corpus that were spontaneously uttered by the subjects (cf. Table 3.3). We classify all these commands in two broad groups: (i) commands that contain focus instances and (ii) commands that do not contain focus instances (e.g., undo commands, redo commands, stop commands, help commands, etc.).

3.5.1 Commands that Contain Focus Instances

Let g_c be the node representing the current focus of attention and let C be a command that comprises following focus instances $f_1, f_2, \ldots, f_n$, where: $n \geq 1$, all focus instances belong to different focus classes, f_1 is the most general focus instance and f_n is the most specific focus instance in the command C. As mentioned above, mapping of a command onto the focus tree is performed with respect to the position of the current focus of attention. The underlying idea could be summarized as follows. In the first step, a temporary focus of attention is positioned on a node that represents the most general focus instance from the command C, i.e., f_1. There can be more than one node satisfying this condition. Thus, the selection of one of them is determined by the position of the node g_c, as discussed below. In succeeding steps, a temporary focus of attention is iteratively transited over nodes that represent focus instances $f_2, f_3, \ldots, f_n$, following the rule that, for all i, j, where $1 \leq i < j \leq n$, the node representing the focus instance f_j is a descendant of the node representing focus instance f_i. The new focus of attention is placed on the node representing the most specific focus instance from the command C, i.e., f_n.

In each of these steps there might be more candidate nodes for a temporary focus of attention. Generally, for a given current focus of attention, a command C can be mapped to different sets of nodes in the focus tree. The transition of a temporary focus of attention may branch with each focus instance from the command C, consequently resulting in more candidate nodes for the new focus of attention. One of these candidates is to be selected to represent the new focus of attention. It is a matter of dialogue context and applied dialogue strategy which candidate node will be selected. The dialogue strategy applied by the dialogue management module in the NIMI-TEK prototype system is discussed in Chapter 4 in more detail. Here, we abstract away from aspects of the dialogue strategy that are not important for the discussion in this chapter and state just the rule for selection among candidate nodes: the first node that was inserted into the set of candidate nodes for the new focus of attention is selected to represent the new focus of attention. Reasons for this rather trivial selection rule will become more clear in Chapter 4. Here we use this rule "as is" to illustrate algorithms for transition of focus of attention.

In order to describe transition of the focus of attention in more detail, we distinguish between two cases. The first case is when each focus instance from the command C can be represented by some of the descendant nodes of the node g_c representing the current focus of attention, i.e.:

$$\{f_1, f_2, \ldots, f_n\} \subseteq descendant(g_c) \tag{3.1}$$

The mapping of the command C is restricted to the sub-tree determined by the node g_c as its root node. In the first step, candidate nodes for representing the most general focus instance f_1 are selected only among descendant nodes of the node g_c. Other nodes that could also represent this focus instance are not taken into consideration in this case. Since selection of a new temporary focus of attention in succeeding steps is always limited to the sub-tree determined by the current temporary focus, final candidate nodes for the new focus of attention are also selected among descendant nodes of the node g_c.

The second case is when not all focus instances from the command C can be represented by some of the descendant nodes of the node g_c, i.e.:

$$\{f_1, f_2, \ldots, f_n\} \not\subseteq descendant(g_c) \tag{3.2}$$

In this case, a temporary focus of attention is first placed on the *closest* antecedent node g_{temp} of the node g_c, that satisfies the condition that each focus instance from the command C can be represented by some of its descendant nodes, i.e:

$$\begin{gathered} g_{temp} \in antecedent(g_c) \\ \wedge R(g_{temp}) = max(R(g_i) | g_i \in antecedent(g_c) \\ \wedge \{f_1, f_2, \ldots, f_n\} \subseteq descendant(g_i)) \end{gathered} \tag{3.3}$$

The command C is then mapped within the sub-tree determined by the node g_{temp}, as described in the first case. Both these cases are encapsulated in the recursive algorithm given in Figure 3.8.

```
procedure select_focus_candidates($g_{arg}$, $f_i$)
begin
  if $\{f_i, \ldots, f_n\} \subseteq$ descendant($g_{arg}$) then
  begin
    S := $\{g | g \in$descendant($g_{arg}$) $\wedge\, g$ represents $f_i\}$;
    for each $g \in$ S do
      if (i = n) then focus_candidates.add($g$)
      else select_focus_candidates($g$, $f_{i+1}$)
  end else
    if ($f_i$ = $f_1$) $\wedge$ (R($g_{arg}$) > 0) then
      select_focus_candidates(parent($g_{arg}$), $f_i$)
    else Exit();
end
```

Figure 3.8: Algorithm for identifying candidate nodes for the new focus of attention.

According to the description above, the first call of this algorithm is to be realized with the following arguments: the node representing current focus of attention g_c, and the most general focus instance from the command C, i.e., f_1, as showed in Figure 3.9. After all recursive calls of this algorithm are finished, candidate nodes for the new focus of attention are accumulated in the set variable `focus_candidates`. If the set of candidate nodes is not empty, one of them is selected (e.g.—for the purpose of illustration and without loss of generality—the first node that has been inserted into the set `focus_candidates`) to represent the new focus of attention. In contrast, the empty set signals that the command C is semantically irregular. Generally, a semantically irregular command includes at least two focus instances f_i and f_j for which holds: $i \leq j$ and f_j is not a sub-focus of f_i, e.g., "translate clockwise". Both the selection among candidate nodes and the system's reaction to a semantically irregular command are to be performed according to the applied dialogue strategy. We address this issue in Chapter 4. The purpose of the variable `history` is explained below.

Let us illustrate these algorithms for the following sequence of commands:

> (C_1:) triangle to right ... (C_2:) now to right rotate ... (C_3:) to left ... (C_4:) upwards ...

At the beginning of this sequence, the current focus of attention is placed

```
procedure change_focus()
begin
  focus_candidates.empty();
  select_focus_candidates($g_c$, $f_1$)
  if (not empty(focus_candidates)) then
  begin
    history.push($g_c$);
    $g_c$ := choose_from(focus_candidates)
  end else irregular_command()
end
```

Figure 3.9: Transition of the focus of attention.

on the root node of the focus tree. Relevant parts of the focus tree are represented in Figure 3.10 for commands C_1 and C_2, and in Figure 3.11 for commands C_3 and C_4. Changes of a temporary focus of attention are marked with dashed arrows. Nodes representing the temporary focus of attention during the mapping of a command are positioned in ovals, while nodes representing the new focus of attention after a command has been mapped are positioned in boxes.

Command C_1 contains two focus instances: f_1 = "triangle" and f_2 = "to right". In the given focus tree, focus instance f_1 can be represented only by the node $\triangle$, while focus instance f_2 can be represented by four different nodes $\{\rightarrow, \curvearrowright, \Rightarrow, \circlearrowright\}$ (cf. Table 3.4). However, for the starting focus of attention placed on the root node, the condition (3.1) introduced above is satisfied, i.e., all focus instance from command C_1 can be represented by some of the descendant nodes of the node representing the current focus of attention. Therefore, all changes of the temporary focus of attention are directed towards more specific focus instances. In the first iteration, when focus instance f_1 is being mapped, the temporary focus of attention is placed on the node $\triangle$. Therefore, in the second iteration, mapping of the focus instance f_2 is restricted to the sub-tree determined by the node $\triangle$ as its root node. There are just two nodes in this sub-tree that are candidates to represent focus instance f_2: $\{\rightarrow, \curvearrowright\}$. Since there are no more focus instances in command C_1 to be mapped, one of these nodes should be selected to represent the new focus of attention. According to the selection rule that we introduced above, the first node → is selected to represent the new focus of attention (cf. left part of Figure 3.10).

Command C_2 contains two focus instances: f_3 = "to right" and f_4

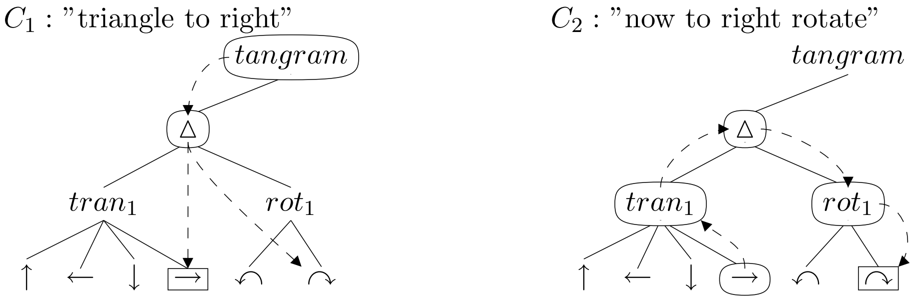
= "rotate". In the focus tree, focus instance f_3 can be represented by nodes $\{\rightarrow, \curvearrowright, \Rightarrow, \circlearrowright\}$, while focus instance f_4 can be represented by nodes $\{rot_1, rot_2\}$. We make two remarks. First, these focus instance cannot be mapped immediately. It is important to note that—keeping in mind that the current focus of attention is placed on the node $\rightarrow$ after command C_1 has been processed—command C_2 satisfies condition (3.2), i.e., not all focus instances from the command C_2 can be represented by some of the descendant nodes of the node $\rightarrow$. Thus, the temporary focus of attention should be iteratively moved towards higher levels of the focus tree until we reach a node whose descendant nodes can represent all focus instances from the command, i.e., a node that satisfies condition (3.3). So, the temporary focus of attention is first placed on the parent node of the node representing the current focus of attention—the node $tran_1$. Since condition (3.3) is still not satisfied, the temporary focus of attention is moved one level higher in the focus tree and placed on the node $\vartriangle$. Now, when condition (3.3) is satisfied, focus instances f_3 and f_4 can be mapped within the sub-tree determined by the node $\vartriangle$ in a similar way as focus instances in command C_1 have been mapped. The second remark is that although focus instance f_3 comes before focus instance f_4 in the utterance, focus instance f_4 is first mapped because it is more general. When focus instance f_4 is being mapped, the temporary focus of attention is placed on the node rot_1. When focus instance f_3 is being mapped, the temporary focus of attention is placed on the node $\curvearrowright$. Since now all focus instances contained in command C_2 have been mapped, the new focus of attention is placed on this node (cf. right part of Figure 3.10).

Mapping of commands C_3 and C_4 is performed in the same way as command C_2. Transition of the focus of attention is illustrated in Figure 3.11.

3.5.2 Commands that do not Contain Focus Instances

This subsection introduces rules for transition of the focus of attention for commands that do not contain focus instances. These commands are classified in six groups (cf. Table 3.3).

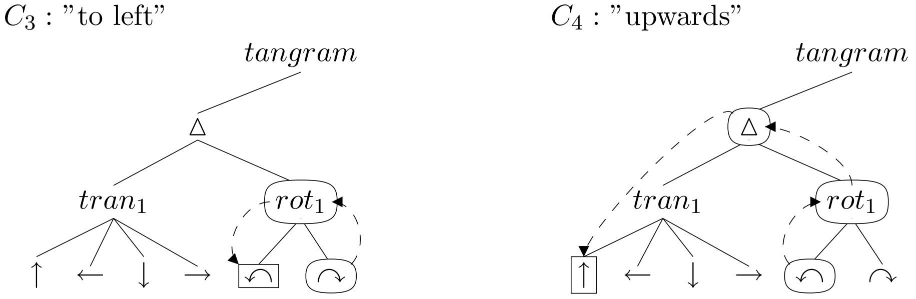
Undo commands. These commands (e.g., "back", "return", etc.) undo a previously performed action. They also restore the previous focus of attention. Since these commands may be successively uttered more than one time, it is necessary to keep track of nodes that represented the focus of attention during the interaction. This is achieved by the variable `history` (see Figure 3.9). As a general rule, every time when a node in the focus tree loses the focus of attention, its unique marker is pushed on the `history`



C_1 : "triangle to right"

C_2 : "now to right rotate"

Figure 3.10: Transition of the focus of attention for the commands C_1 and C_2.



C_3 : "to left"

C_4 : "upwards"

Figure 3.11: Transition of the focus of attention for the commands C_3 and C_4.

stack. Thus, when an undo command is uttered, the new focus of attention is assigned to a node whose marker is popped from the `history` stack.

Redo commands, ***Stop commands*** and ***Help commands***. Redo commands explicitly instruct a repetition of a previously performed action (e.g., "further", "more", "again", etc.). Stop commands explicitly instruct an immediate termination of a currently performed action (e.g., "stop", "don't rotate", etc.) In Help commands, the user explicitly asks for support. A common point for all these commands is that they do not change the current focus of attention. When Redo commands are uttered, only the unique marker of the node in the focus tree representing the current focus of attention is pushed on the `history` stack.

Commands specifying a focus-class. Commands from this group specify only a focus class, but not a focus instance (e.g., "give me a triangle", etc.). In effect, the subject allows the system to choose among instances from the specified focus class. The actual decision of a system, however, is a matter of dialogue context and of the applied dialogue strategy that is introduced in Chapter 4.

Commands containing ellipsis-substitutions. In these commands, the subjects used ellipsis-substitutions (e.g., "do what i say", etc.) to signal that the system's performance was incorrect (cf. Appendix A). However, these commands do not provide additional information about the required focus transition. The system may ask for a clarification, provide support or wait for a succeeding command, which is, again, a matter of the applied dialogue strategy.

3.6 Advantages and limitations

In this section, we discuss advantages and limitations of the proposed approach to processing of the user's commands. We make several remarks.

(1) Phrasal lexicon. The first set of remarks is on phrasal lexicon. Illustrating the introduced algorithms for transition of the focus of attention, we observed focus instances that are contained in a command (e.g., command "triangle to right" contains two focus instances: "triangle" and "to right", etc.), but we did not explain how these focus instances were extracted from a command. Here, we shortly discuss this issue. To each focus instance in the focus tree a set of phrases that represent it is assigned. In processing users' commands, the system takes as input a textual version of the user's command outputted from the speech recognizer. The focus instances are then automatically derived from a given command, i.e., the system detects phrases that relate to certain focus instances. We took phrases from the NIMITEK corpus uttered by the subjects and correlated them with the corresponding focus instances. For example, the focus instance that is represented by the node □ in the focus tree may be correlated with the following phrases {square, yellow square, quadrangle, ...} (in German: {Quadrat, gelbes Viereck, Viereck, ...}). In addition, sets of phrases assigned to different focus instances are not necessary disjoint sets. For example, explaining how command C_1="triangle to right" was processed (cf. Subsection 3.5.1, Figure 3.10), we stated that the phrase "to right" could be assigned to four different nodes in the given focus tree $\{\rightarrow, \curvearrowright, \Rightarrow, \circlearrowright\}$ and illustrated how the introduced algorithms cope with such a situation.

Context dependent commands are also taken into account. These commands contain phrases that cannot be uniquely related to a focus instance in the focus tree, e.g., "back", "further", "the next triangle", etc. (in

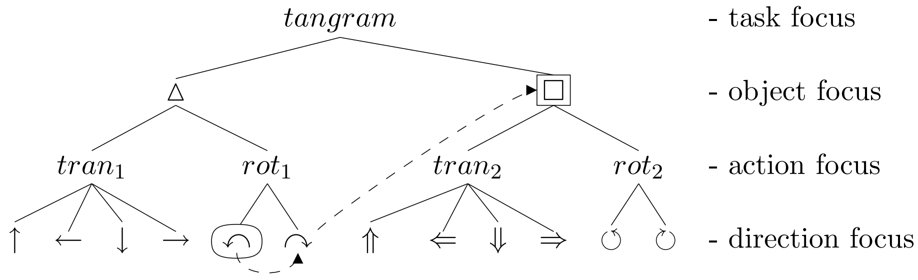


Figure 3.12: Transition of the focus of attention for the context dependent command C_5.

German: ”zurück”, ”weiter”, ”nächstes Dreieck”, etc.). We differentiate among several types of context dependent commands: Undo commands, Redo commands, Stop commands, Help commands, commands specifying a focus class, etc. It has been already discussed in Subsection 3.5.2 how is the focus of attention managed for these commands. Generally, the history of interaction[3] and the structure of the focus tree enable the system to process such commands. Here we provide a small example that illustrates how is the context dependent command C_5=”the next piece” processed. This commands specifies only a focus class (i.e., object focus), but not a focus instance. In other words, a Tangram piece should be selected, but it is not explicitly specified which Tangram piece should be selected. The processing of this command can be summarized in the following steps: We traverse the focus tree in preorder, starting from the node that represents the current focus of attention. The new focus of attention is placed on the first node that satisfies the following conditions: (i) it represents a focus instance that belongs to the object focus class, and (ii) the selected node is not the starting node. This is illustrated in Figure 3.12. The node representing the starting focus of attention is positioned in oval, the node representing the new focus of attention is positioned in box, and the relevant part of the preorder traversal is marked with dashed arrows.

(2) Generalizability. The second set of remarks is related to the issue of generalizability. The processing of commands were illustrated for the Tangram puzzle. A question that arises is to what extent can this approach be

[3] The history of interaction, that is also a part of our implementation (see Chapter 4, Subsection 4.4.5), should not be confused with the `history` stack introduced in Figure 3.9—this variable is only a part of the more general structure that represents the history of interaction.

generalized. We discuss this question from two points of view: the engineering point of view and the linguistic point of view (cf. Allen 2008).

The engineering point of view considers primarily implementation aspects. The proposed modeling method and algorithms are not *a priori* related to some specific predefined task. The introduced algorithms are independent of the structure of the focus tree and of the content of the phrasal lexicon. For a given task (e.g., Tangram puzzle, Tower of Hanoi puzzle, etc.), the structure of the focus tree and the sets of phrases that are assigned to focus instances are defined in input XML files, independently of the implementation of the algorithms introduced in this chapter. This means that the implementation of the proposed model of attentional state within the dialogue management module in the NIMITEK prototype system is independent of:

- changes of the structure of the focus tree (e.g., a change from the Tangram puzzle to the Tower of Hanoi puzzle, etc.),
- changes of the vocabulary (e.g., changing the size of the vocabulary by extending or redefining sets of phrases, changing the language of the vocabulary by translating phrases from German into English, etc.).

These changes do not require a change in the core implementation, but just a redefinition of input XML files. From the engineering point of view, this gives a relatively high level of generalizability of the proposed model—the given task can be relatively easy redefined or extended. This is discussed and illustrated in Chapter 5 in more detail.

The linguistic point of view considers the question to which types of dialogue can this approach be applied. It is clear that it does not cover a general case of unrestricted dialogue. But it was not our intention in the first place. We concentrate on spoken human-machine interaction in the specific case where some kind of display with a graphical interface is involved. In such cases, all "relevant" focus instances may be known in advance as they are also a part of the graphical interface, and, in addition, we can define subfocus relations between them (i.e., we can define a focus tree, cf. discussion in Section 3.4). These two requirements are needed to apply the proposed modeling methodology. In other words, it can be summarized that our approach to modeling attentional information is appropriate for the class of spoken dialogue systems that are intended to control a subclass of graphical user interfaces, e.g., manipulating with graphical entities represented on the display, controlling graphical menus, solving graphically-based tasks and playing interactive board games that includes spatial reasoning, etc.

The second limitation is that the proposed approach supports processing of commands uttered by users, but it does not support processing of questions and statements that could also relate to the focus of attention. In addition, it also does not support dialogue acts such as (cf. Allen 2008) indirect speech acts (e.g., "How much time do I still have?", etc.) and meta-discussion (e.g., "I don't feel like doing", etc.). However, these classes of dialogue acts are clearly less frequent than commands in the NIMITEK corpus.

Another limitation is the lack of compositionality. We provide two small illustrations. As the first illustration we consider the command "move the triangle and the square rightwards". This command contains two focus instances "triangle" and "square" that belong to the object focus class. Therefore, it is necessary to divide this command in a sequence of commands whose all focus instances belong to different focus classes (e.g., "move triangle rightwards" and "move square rightwards") before it can be processed. The second illustration relates to phrasal lexicon. We stated that to each focus instance in the focus tree a set of phrases that represent it is assigned. One of the phrases that can be assigned to the focus instance △ in the focus tree that we observe in this chapter is "triangle" (in German: "Dreieck"). Inspection of the commands from the NIMITEK corpus shows that subjects used the noun "triangle" 224 times (mostly in nominal phrases) in order to instruct that a triangle Tangram piece should be selected. Some examples are: "the small triangle" (in German: "das kleine Dreieck"), "yellow triangle" (in German: "gelbes Dreieck"), "the left triangle" (in German: "das linke Dreieck"), etc. However, in 10 commands the noun "triangle" was used either to explicitly instruct that another Tangram piece (i.e., not a triangle) should be selected, or to specify an action that is to be performed over previously selected Tangram piece (that is not necessarily a triangle). Examples of these commands are: "the shape on the right side that is not a triangle" (in German: "der ganz rechte Körper, der kein Dreieck ist"), "please place on the triangle" (in German: "Bitte platzieren Sie auf das Dreieck"), etc.

Although such "problematic" commands are rare in the NIMITEK corpus, compositionality is in general an important issue that is not to be left disregarded. Therefore, it is important to note that lack of compositionality is not inherent to the proposed modeling approach. In other words, the issue of compositionality may be addressed independently of the proposed approach to modeling attentional information. We explain this shortly. Here, the issue of compositionality is related to the research question of automatic extraction of focus instances from users' commands. Various approaches may be applied for this purpose, starting from simple string comparison to more complex grammar frameworks. This research question is not considered in this chapter (although the author is aware of its importance). Another research question is how extracted focus instances may be used to process users' commands of different syntactic forms. This research question is considered in this chapter. Furthermore, these two research questions may be addressed independently, i.e., extraction of focus instances is independent of the introduced algorithms for transition of the focus of attention. Once focus instances have been extracted from a command, they are delivered as input to the introduced algorithms.

Finally, we point out that our approach to modeling attentional information is not limited only to verbally uttered commands. It supports also non-verbal dialogue acts produced by the user (e.g., using a mouse or a keyboard, etc.) or by the system (e.g., performing a move). Such non-verbal actions may change also the attentional state. For example, if the user was allowed to control a mouse to "click" on a graphical piece represented on the screen (e.g., the square) in order to select it, she would thereby unambiguously specify that the current focus of attention should be placed on the node in the focus tree that represents the selected piece (i.e., in our example, the node □). We recall that the NIMITEK prototype system was not designed to accept user input from keyboard or mouse. However, it is simply a specification requirement, and not a deficiency of the underlying dialogue model. It is clear from the discussion that extending the system to accept user input from keyboard or mouse (or touch-screen, etc.) would not be problematic with respect to modeling of attentional state. On the other hand, the NIMITEK prototype system produces non-verbal dialogue acts frequently (e.g., performing users' commands). To illustrate this, let us consider a situation when the user has problems to solve a given task. The system may, as a part of support, propose and perform a move. For example, the system may select a graphical piece to be moved or move a previously selected piece on a target position, etc. Again, it is unambiguously determined how the focus of attention should be changed. Transition of the focus of attention for non-verbal dialogue acts produced by the system is—although rather trivial to implement—important for processing of subsequent user's commands (e.g., when the user instructs an "undo" command immediately after the system proposed and performed its move, etc.). This is illustrated in Chapter 5.

3.7 Conclusion

In this chapter we proposed an approach to processing of the user's commands in human-machine interaction for the restricted model of commands contained in the NIMITEK corpus. Based on general observations on the structure of spoken language made by Campbell (2006), on the theory of discourse structure introduced by Grosz and Sidner (1986) and, finally, on the inspection of the domain-specific NIMITEK corpus, we introduced (1) the concept of the focus tree in order to model attentional information on the level of the user's command and (2) the rules for transition of the focus of attention for different types of commands. The processing of commands were illustrated for the Tangram puzzle. We also discussed advantages and limitations of the proposed approach, including also the question to what extent can this approach be generalized.

Still, this chapter did not consider all important implementation aspects of the model of attentional state in the dialogue management module of the NIMITEK prototype system. In addition to processing of users' commands, we use the focus of attention as one of interaction features to model the state of the interaction. Thus, several important questions remained to be discussed. Some of them are: How is the focus tree to be used in the framework of an adaptive dialogue strategy aimed to support the user to overcome problems in interaction? How do the proposed algorithms work in a realistic scenario? How this model may contribute to overcome some limitations of automatic speech recognition technology? These questions are discussed in the coming chapters.

Chapter 4

An Adaptive Dialogue Strategy

4.1 Introduction

In the previous chapter, we discussed some aspects of natural language interfaces. However, the issue of naturalness of human-machine interaction considers more than just the language interface. While in the previous chapter we primarily considered the question how to understand the user's utterances (e.g., to understand which move was instructed by the user while she solves a graphical puzzle), here we consider implications that understanding of the user's utterances may have on the dialogue situation (e.g., how does the user's move change the state of the puzzle, whether it is useful or legitimate, does it draw back the state of the puzzle from the expected final state, etc.).

In addition to the language interface, we state two requirements that we find to be essential in achieving a higher level of naturalness of the interaction:

- The behavior of the system should be dynamically adapted according to the current state of the interaction.
- The user should be considered as an integral part of the interaction. Consequently, providing a response, systems should also take properties of the user—especially the emotional state of the user—into account.

In this chapter, we discuss both these points. We already motivated the need for dialogue strategies to support the user to overcome problems that occur in the interaction (cf. Chapter 1, Section 1.1). The aim of this chapter it to propose an approach to designing adaptive dialogue strategies. More precisely, this chapter reports about design and implementation of the adaptive dialogue strategy in the NIMITEK prototype spoken dialogue system for supporting users while they solve a problem in a graphics system. As it is clear that this approach is not intended to cover the general case of unrestricted human-machine interaction within arbitrary domains, it should also not be understood that this approach is limited to the interaction domain of the NIMITEK prototype system only. It covers the class of spoken dialogue systems that are intended to manage a subclass of task-oriented dialogues, i.e., dialogues that are primarily concentrated on a given task, where the state of the task[1] is observable in the sense that it can be explicitly defined and evaluated with respect to how it corresponds to expected final states. The underlying idea is that the system dynamically adapts its dialogue strategy according to the actual state of the interaction. In order to make the system able to select and apply an appropriate adaptation of the dialogue strategy, various interaction features should be taken into account. We call the composition of these features *the state of the interaction.* For the purpose of this contribution, we consider five interaction features: the state of the task, the focus of attention (introduced in the previous chapter), the user's command, the state of the user, and the history of interaction.

These issues are considered in more detail below. Section 4.2 provides an overview of the state-of-the-art in dialogue management, focusing particularly on adaptive dialogue systems and dialogue strategies that take emotional state of the user or the state of the interaction into account. In Sections 4.3 and 4.4, we motivate and introduce a model of the state of the interaction that includes above mentioned interaction features as integral parts. In Section 4.5, we report about design and implementation of an adaptive dialogue strategy for supporting the user to overcome problems in the interaction. Section 4.6 describes briefly the functionality of the dialogue management module and its relations to functionalities of other modules incorporated in the NIMITEK prototype system.

[1] The state of the task should not be confused with the state of the interaction. As we explain in this chapter, the state of the task is only a part of the more general structure that represents the state of the interaction.

4.2 Background and Related Work

In the last fifty years, many approaches to dialogue management in human-machine interaction have been introduced and implemented, both in academic and industry fields, covering various domains of interaction: providing timetable information, tickets booking, interactive planning, tutoring, medical advising, interactive data analyzing, etc. In this section, we mention[2] some of the most important dialogue systems and shortly state main aspects of the used approaches to dialogue management.

The SUNDIAL (Speech UNderstanding in DIALogue, cf. Bilange 2000; Catizone et al. 2002) system is designed to maintain dialogues with users over standard telephone lines. The domain of this system is flight reservations and enquiries, and train enquiries. With respect to dialogue management, this project integrates two lines of research. The first line is that dialogue management should be generic, i.e., not oriented to one task domain only or to one interface language only. The second line of research is that dialogue management should maintain a natural and cooperative interaction with the user. Output produced by the system *should be perceived as natural, coherent, and helpful within the context of the dialogue* (Catizone et al., 2002, p. 12). In addition, untrained, linguistically naive users should be allowed to express themselves naturally, and the dialogue manager should be capable to deal with problems caused by non-accurate speech recognition. To achieve these two requirements in the SUNDIAL system, an interactional model is introduced. It includes four structural layers: a linguistic structure, a belief structure , a dialogue structure, and a task structure. The SUNDIAL system is based on a distributed architecture, where each layer of the interactional model is represented in a separate module. All modules communicate with each other to provide or to get the relevant information. The system supports interaction in four languages: French, German, Italian and English.

The ARISE (Automatic Railway Information System for Europe, cf. Catizone et al. 2002) system is a spoken dialogue system for providing train timetable information over telephone lines. This system is based on modular architecture, including the module for mixed-initiative dialogue management. In order to increase the level of naturalness of the interaction and, in

[2] We do not intent to provide a complete overview of the state-of-the-art in dialogue systems, but only to discuss researches that we find particularly interesting for the discussion in this chapter. The more complete overviews of the state-of-the-art in dialogue system are already given in various reports (cf. Wilks et al. 2006, Catizone et al. 2002, Xu et al. 2002).

the same time, to support ”correct” development of the interaction, the dialogue manager combines two approaches to maintaining the dialogue: a flexible mixed-initiative approach when the system is confident that the user’s utterance has been correctly understood, and a restricted system-directed approach when the system detects problems in communication. The ARISE system supports interaction in four languages: Dutch, French, English, and Italian.

The SmartKom (cf. Alexandersson and Becker 2001; Catizone et al. 2002) is a multimodal dialogue system that combines three knowledge resources in maintaining the interaction with the user: speech, gestures and facial expressions. One of the application scenarios for this system is an intelligent telephone booth that provides an interaction interface for booking tickets, providing information about different local activities and attractions, etc. The architecture of the SmartKom includes—among other modules—semantic processing modules for gesture and speech analysis, media fusion, intention recognition, discourse and domain modeling, action planning, presentation planning and concept-to-speech generation (Catizone et al., 2002, p. 15).

The TRAINS-95 (Allen, 2008; Catizone et al., 2002) is an end-to-end plan-based dialogue system that was developed at the University of Rochester in order to demonstrate that it is possible to achieve robust system’s performance in some problem solving domain. It was aimed to address the main disadvantages of previous plan-based dialogue systems that had lacked reasonable coverage of the domain, and been to slowly and fragile. The domain of this system is route planning. The route planning part of the system was deliberately implemented not to be optimal to enforce interaction between the user and the system. The TRAINS-95 system demonstrated robust understanding and robust performance with untrained subjects using speech on simple route finding problems. Its successor, the TRIPS-98 system (the Rochester Interactive Problem Solver) demonstrated ”mixed-initiative planning” using state of the art planning technology, and integrated dialogue with sophisticated back-end reasoning. In addition, this system handles some complex discourse behavior such as ellipses, clarifications, corrections, etc. However, three main limitations of the TRIPS-98 system remained to be addressed. First, the system acts only in response to user inputs, i.e., although it implements mixed-initiative planning, it does not support mixed-initiative dialogue. Second, there is no ”meta-discussion” capability, e.g., it does not support dialogue acts in which the user tries to shift the topic of the dialogue, etc. Third, the implementation of the system was limited to the functionality of planning only. The next successor system is a mixed-initiative medical advisor. In the architecture of this system, the agency of the system is separated from dialogue management. The first advantage of this architecture is that mixed-initiative is implemented in a principled, not domain oriented, way. The second advantage is that collaborative problem solving acts (e.g., adopt, defer, evaluate, identify) provide an abstract interface to back-end reasoning system.

The descriptions of these systems underline two long-term research goals: to increase the level of naturalness of the interaction, and to increase probability that a dialogue will be successfully concluded when problems occur in the interaction. To achieve these goals, in the last decade we witness the rapid increase of research interest in adaptive dialogue systems and affected user behavior. In the rest of this section, we mention several researches that we find particularly interesting for the discussion in this chapter. These researches primarily relate to adaptive dialogue systems and dialogue strategies that take emotional state of the user or the state of the interaction into account.

Luzzati (2000, p. 220) proposes a dialogue model that *increases probability that a dialogue will be successfully concluded, even with a limited understanding level*. It is intended to suggest a method of computing dynamically how to choose between dialogue strategies. In his own words, *instead of trying to compute the intentions of the user, this model tries to give the machine a kind of awareness of the state of the interaction* (Luzzati, 2000, p. 209). In an uninterrupted communication, the user and the system exchange questions and answers related to a given communicational topic. Otherwise, when a problem emerges in the interaction, dialogue acts such as reformulation, iteration, restarting, explanation, etc. are used in order to resolve the problem. Luzzati considers relations between users' and the system's dialogue acts in order to model the dialogue state. His model chooses an appropriate dialogue strategy among available strategies (e.g., asking the user to repeat her preceding question, to provide an explanation, to start the whole phase of dialogue again, etc.) according to the actual dialogue state. Although giving an account on the systemic question/answer relationships that are present in the structure of conversation turns out to be a challenging task (cf. Searle 1992a,b), the idea that a dialogue strategy can be dynamically computed with respect to the state of the interaction still remains worth to be further investigated. Building discourse models that may be consulted by the system in order to select an appropriate dialogue strategy is a popular approach to handle miscommunication in the context of spoken natural language human-machine interaction (cf. McTear 2008; McTear et al. 2005, Skantze 2008b,a). This is discussed in more detail in

Chapter 5 (Section 5.4.1). However, it should be noted that Luzzati does not consider the state of the user as an integral part of the state of the interaction.

Heckmann et al. (2007, 2005) introduce the general user model ontology (GUMO) for the uniform interpretation of distributed user models in intelligent semantic web enriched environments. The major advantage of the GUMO is *the simplification for exchanging user model data between different user-adaptive systems* (Heckmann et al., 2007, p. 37). This ontology includes different user's dimensions that are modeled within user-adaptive systems. Some of them are: *emotional state* (with sub-types: happiness, anxiety, pride, shame, satisfaction, confusion, etc.), *characteristics* (with sub-types: talkative, assertive, dominant, quiet, kind, etc.), *personality* (with sub-types: extravert, introvert, thinking, judging, optimistic, etc.), *physiological state* (with sub-types: heartbeat, blood pressure, respiration, temperature, injury, etc.). An interesting point is that—in accordance with the intention to introduce an ontology that is suitable for user-adaptive systems—dimensions carry *the qualitative time span of how long the statement is expected to be valid (like minutes, hours, days, years)* (Heckmann et al., 2007, p. 40). For example, heartbeat can be changed within seconds, emotional state can be changed within minutes, personality can be changed within months or years, and demographics information such as birthplace is not expected to be changed at all. The state of the user carries important information for managing the affected user behavior. Still, we make two remarks. First, Heckmann et al. do not consider dialogue strategies that take the emotional state of the user into account. And second, their encoding of emotional states is just one of many different ways to encode emotion-related content. Different emotions occur in different application scenarios. Recognition and interpretation of emotions as well as their level of significance may vary with respect to a given scenario. We discuss this in more detail in Subsection 4.4.4.

Batliner et al. (2000) recognize the significance of detecting emotional user behavior for successful automatic dialogue processing. In the framework of the Verbmobil project, they report about the module *MOUSE* (*Monitoring of User State [especially of] Emotion*, Batliner et al. 2000, p. 198–200) that combines several knowledge resources (e.g., the classifier for prosody, the classifier for repetitions and reformulations, etc.) within an integrated classification of trouble in communication. This module is not fully integrated in the Verbmobil system but can be switched on for demonstration purposes (Batliner et al., 2000, p. 195). Their point of departure is that *the user behavior is supposed to mirror the state of the communication* (Batliner et al., 2000, p. 198). If there are no troubles in the communication, the user behaves neutral and is not engaged emotionally. Otherwise, the user behavior changes accordingly (e.g., overt signaling of emotions, etc.). Batliner et al. are primarily concentrated on the problem of timely recognition of troubles in communication. Although they are aware of the need to develop appropriate dialogue strategies for resolving these troubles, they do not consider this issue in more detail.

An example of adaptive spoken dialogue system intended to recognize and respond to emotion of the user is ITSPOKE (Intelligent Tutoring SPOKEn dialogue system), reported by Litman and Silliman (2004) and Forbes-Riley et al. (2008b). It is a spoken dialogue system that uses an already existing text-based, conceptual physics tutoring system as back-end. ITSPOKE has been built to assess the impact and evaluate the utility of adding spoken language capabilities to dialogue tutoring systems. The point of departure was the assumption that speech-based tutorial dialogue systems could be more effective then text-based systems. This assumption was based on observations that spontaneous self-explanation by students improves learning gains during human-human tutoring, and that spontaneous self-explanation occurs more frequently in spoken tutoring than in text-based tutoring. The interaction between the student and the system can be summarized as follows: ITSPOKE poses a conceptual physics problem in textual form and the student types in and submit a natural language essay answer, after which the spoken dialogue between the student and the system begins. The spoken dialogues have a Question-Answer-Response format, implemented with a finite state dialogue manager. Responses of the system depend on the correctness of the student's answer and the difficulty of the question. If the answer is correct, the system moves on to the next question. For incorrect answers to easier questions, the system provides the correct answer accompanied with a brief statement of reasoning. For incorrect answers to more difficult questions, the system engages the student in *a remediation subdialogue, containing questions that walk the student through the more complex line of reasoning required for the correct answer* (Forbes-Riley et al., 2008b, p. 61).

In addition, Litman and Silliman (2004, p. 5) initially hypothesized that *the success of computer tutors could be increased by recognizing and responding to student emotion.* Thus, ITSPOKE has been enhanced also to automatically respond to student affect. Considering student affect (i.e., emotions and attitudes), Forbes-Riley et al. (2008b) target—uncertainty. The first reason for focusing on uncertainty is that it occurred more than other affective states in prior ITSPOKE dialogues. The second reason is based on the observation that uncertainty and incorrectness can be seen as signalling *learning impasses: opportunities for the student to learn the material about which s/he is uncertain or incorrect.* Forbes-Riley et al. (2008b, p. 61) restricted the initial hypothesis in its scope: *Responding to uncertainty in the same way as incorrectness will improve student performance, by providing students with the knowledge needed to resolve their uncertainty impasses.* An initial investigation of the impact of this adaptation on student performance is conducted in a Wizard-of-Oz simulation of the ITSPOKE. Three conditions were used in this experiment to test whether the uncertainty adaptation improves student performance:

- experimental condition—the dialogue manager adapts to uncertainty by treating all uncertain+correct student answers as incorrect,
- normal control condition—the dialogue manager does not adapt to uncertainty,
- random control condition—the dialogue manager does not respond to uncertainty, but it treats a percentage of random correct answers as incorrect.

Results of this experiment—though not conclusive—suggest that the uncertainty adaptation does have a positive benefit on student performance (Forbes-Riley et al., 2008b, p. 68).

Another interesting example is the pedagogical agent introduced by Burleson (2006). His research focuses on the role of feelings in learning:

> It uses an affect-focused approach to supporting learners by sensing their level of frustration and assessing the appropriateness of task based vs. affect based interventions. (Burleson, 2006, p. 12)

Based on theories of metacognition that describe how learners use strategies and self-awareness to improve their thinking process, Burleson (2006, p. 13) defines *meta-affect* as comprising three things:

- Meta-affective knowledge: knowledge about how affect works, e.g., a failure to accomplish a given task could induce frustration.
- Meta-affective experience: online awareness of feelings and the user's conscious reflection on what that emotion is doing to her/him, e.g., frustration is the reason why the user wants to quit.

- Meta-affective skill: the ability to coordinate meta-affective knowledge and meta-affective experience.

Burleson (2006, p. 19–22) emphasizes the significance of engagement of the user in a learning process. In order to keep a frustrated user engaged, many Intelligent Tutoring Systems simplify a given task. Burleson argues that such an approach is not always appropriate, e.g., the nature of the task may be unsuitable for modification. In addition, it does not support the user to learn affective self-awareness. He proposes an affective approach based on two premises: awareness of one's affective state can influence a person's ability to alter that state; and pedagogical agents can help learners to realize that they can use a feeling of frustration as a signal that it may be time to try a different strategy. Therefore, his Intelligent Tutoring System facilitates learners' development of both metacognitive and meta-affective skills.

Burleson (2006, p. 14–15) uses the Affective Agent Research Platform that contains a set of sensors to detect the user's emotions, e.g., the pressure mouse, the skin conductance sensor, the seat posture chair, the facial expression camera, etc. The platform contains also a real-time scriptable character agent that is capable of a wide range of expressive interactions with the user. The system's sensors and dynamic scripts enable the pedagogical agent to involve in non-verbal social mirroring of the user's emotional state. In this way, the system helps the user to become aware of her emotional state and to use this awareness for a positive change of a dialogue strategy.

Subjects from the sampled population of 11–13 year-old children participated in an experimental evaluation of the system. They were engaged in the Tower of Hanoi activity—the same domain as in the NIMITEK prototype system. There were four experimental contrasts derived from a 2x2 design (Burleson, 2006, p. 47):

- interactions between the system and the child were either *sensor-driven non-verbal mirroring* or *prerecorded non-verbal interactions*,

- interventions by the system were either an *affect support intervention*, designed to attend to the emotional state of the learner or *task support intervention*, designed to provide the learner with constructive information about the activity.

Making a remark that the generalization of the experimental results to a broader population should be additionally verified, Burleson (2006, p. 79–89) reports:

- There was no support for the hypotheses (a) that the affective learning companion will be more persuasive and that users will form a stronger social bond based on mirroring, (b) that social bond would positively correlate with perseverance or with self-theories, and (c) that an affective learning companion that exhibits emotional intelligence will increase learners' intrinsic motivation and reduce frustration.

- There is minimal to mixed support for the hypotheses (d) that persuasion will positively correlate with social bond and perseverance, and will negatively correlate with frustration, (e) that metacognitive/meta-affective skill will positively correlate with perseverance, willingness to continue, and intrinsic motivation.

An application of argumentation technology that supports medical tasks is reported by Mazzotta et al. (2007). They introduce the prototype persuasion system Portia that is the argumentation module of a dialogue system they have developed. The goal of the system's argumentation is to persuade users to adopt more healthy eating habits. Mazzotta et al. (2007, p. 42) note that eating habits *solidify over time and are difficult to modify* and that *attempting to persuade people to adopt more appropriate habits by employing only rational and scientific arguments are probably ineffective.* Therefore, in their approach they combine rational and emotional strategies.

Their approach is based on observations of how people behave when they want to persuade someone to adopt eating habits. In the first experimental study, they produced a corpus of natural language messages collected from the Italian subjects that were playing the role of persuaders. Two versions of the persuasion scenario were presented randomly to the subjects: they were advised to use either positive or negative arguments. The main findings of this study were (Mazzotta et al., 2007, p. 43):

- No matter how the scenario was formulated, the subjects tended to combine negative and positive arguments, but preferred positive arguments to negative ones.

- Few messages were formulated according to a rational scheme: subjects usually combined rational and emotional arguments, with a prevalence of emotional arguments.

- The recommended behavior was usually introduced at the beginning of mostly rational texts, and only subsequently in more emotional ones, after preparing the subject to receive the suggestion.

In the second experimental study, they engaged a new group of subjects from various countries in order to evaluate persuasion strategies from the corpus. The main findings of this study were (Mazzotta et al., 2007, p. 44):

- The subjects considered the positive emotional version of the dialogue the most persuasive on average.

- The negative emotional version of the dialogue raised quite negative comments: subjects saw the scenario as terrible and the persuader as violent.

- Many subjects claimed that suggestions should be more tailored to the target person, less straightforward, and more cautious, and that the persuader should have engaged the target person in the discussion.

The Portia system performs several tasks: acquires information about the user, provides information on request or according to its own plans, suggests lines of action when appropriate, tries to persuade the user to follow them when needed, and enters into an argumentation subdialogue to justify and support its choices or revise them if needed (Mazzotta et al., 2007, p. 49).

4.3 Guidelines to Adaptive Dialogue Management

The overview of the state-of-the-art in dialogue management given above underlines two long-term research goals: to increase the level of naturalness of the interaction, and to increase probability that a dialogue will be successfully concluded when problems occur in the interaction. In the introduction of this chapter, we stated two requirements for spoken dialogue systems, in addition to natural language interfaces discussed in the previous chapter, that we find to be essential in achieving these research goals. First, the behavior of the system should be dynamically adapted according to the current state of the interaction. And second, the user should be considered as an integral part of the interaction. In this section, we motivate these requirements in more detail.

We model the state of the interaction as a composite of five interaction features: the state of the task, the focus of attention (introduced in the previous chapter), the user's command, the state of the user, and the history of interaction. They are illustrated with respect to interpreting the user's commands and providing support to the user.

Interpreting the user's commands. Recalling the settings of the WOZ experiment described in Chapter 2, the subjects had the freedom to formulate their own instructions. Thus, they were expressing propositional content in different ways. That makes interpretation of their commands more challenging. Let us observe, for example, the following utterances taken from the NIMITEK corpus and produced by the subjects while they were solving the Tower of Hanoi puzzle:

1. The second smallest ring on the two (in German: "Den zweitkleinsten Ring auf die Zwei").
2. Number one on number two (in German: "Nummer eins auf Nummer zwei").
3. The next ring on the two (in German: "Den nächsten Ring auf die Zwei").
4. On the two (in German: "Auf die Zwei").
5. Back (in German: "Zurück").

Although all these commands might even have the same propositional content, from the aspect of interpretation they cannot be treated equally. The first command is clear—the middle ring should be moved on the second peg—whereas remaining commands cannot be interpreted without additional information. In the second command, the ring that is on the top of the first peg should be moved on the second peg. Thus, to properly interpret this command, we need to know the state of the task. Also, in the rest of commands we need to know which ring was last recently moved or selected in order to conclude which ring should be moved. More general, we need to know the history of interaction.

Providing support to the user. We differentiate between several kinds of support. One of them relates to the puzzle itself. For example, in a situation when the user does not know how to solve the puzzle, the system might propose the next move. This kind of support is determined by the state of the puzzle. Another kind of support is related to situations when problems occur at the level of the interface language (e.g., the user knows what to instruct, but does not know how to formulate a proper command). The aim of support in this case is not necessarily to provide the user with new information, but to help her to overcome an interface problem. In such situations, the information about the user's command and the current focus of attention may be required to provide appropriate support. Finally, the manner of providing support should be in accordance with the emotional state of the user, e.g., a frustrated or indisposed user should not be treated in the same manner as a user that is engaged in the task, even if the informational content of support might be the same in both cases.

It is important to note that the model of emotional user states is not predefined. Our point of departure is that different emotions may occur in different application scenarios. Recognition and interpretation of emotions as well as their level of significance may vary with respect to a given scenario. Therefore, after evaluating and analyzing the NIMITEK corpus with respect to its emotional content, we introduce a data-driven model of user emotional states specific for the given interaction scenario. This is described in more detail in Subsection 4.4.4.

4.4 The State of the Interaction

One of the ideas underlying the dialogue management module in the NIMITEK prototype system is to give the system a kind of awareness of the state of the interaction. As motivetd above, we model the *state of the interaction* as a composite of five interaction features:

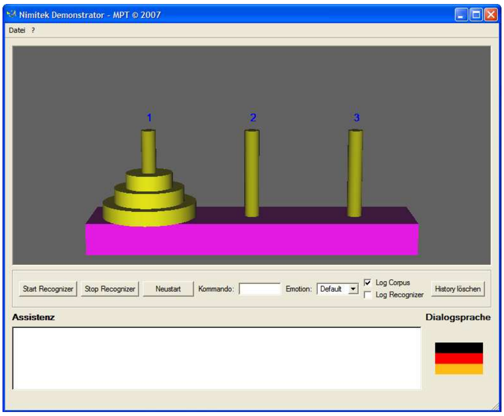
- *the state of the task*,
- *the focus of attention*,
- *the user's command*,
- *the state of the user*,
- *the history of interaction.*

These interaction features may be considered to be of general nature. This chapter primarily addresses the implementation of the dialogue management module incorporated in the NIMITEK prototype system. Thus, for the purpose of better clarity, we describe these interaction features with respect to the existing implementation. However, we note that this implementation-oriented description of the interaction features implies by no means a task-orientation of the proposed approach to designing adaptive dialogue strategies.

4.4.1 The State of the Task

The dedicated prototypical task implemented in the NIMITEK prototype system is the *Tower of Hanoi* puzzle introduced by Édouard Lucas in 1883.

The puzzle consists of three pegs and several disks of different sizes. Our implementation includes three versions of the puzzle: the 2-disks version, the 3-disks version, and the 4-disks version. Without loss of generality, we restrict temporarily our discussion to the 3-disks version of the puzzle. At the start of the game, the disks are stacked in order of size on the leftmost peg, as shown in Figure 4.1. The goal of the puzzle is to move the entire stack to the rightmost peg according to the following rules: only one disk can be moved at a time, only the upper disk from one of the pegs can be moved onto another peg, and no disk may be placed on top of a smaller disk.



Emotion: Default

☑ Log Corpus

☐ Log Recognizer

Figure 4.1: The 3-disks version of the Tower of Hanoi puzzle: Screen display of the NIMITEK prototype system.

The state of the task in the Tower of Hanoi puzzle is defined by the current positions of the disks. For the 3-disks version of the puzzle, the state of the task is represented as an ordered collection of three integer numbers (p_1, p_2, p_3), where $p_i, i \in \{1, 2, 3\}$, is the number of the peg on which the ring i is placed. To illustrate: the starting triple (1, 1, 1) represents the state when all rings are positioned on the first peg, and the triple (3, 1, 1), that usually relates to the state of the task after the first move, represents that the smallest ring is placed on the third peg, while the other rings are still on the first peg. In a version with n disks, the state of the task is represented as an ordered collection of n integer numbers and there are 3^n possible states of the task.

4.4.2 The Focus of Attention

In the previous chapter, we proposed an approach to modeling attentional information on the level of a user's command for the restricted model of commands contained in the NIMITEK corpus. We introduced a tree structure—*the focus tree*—inspired by the concept of the focus space stack suggested by Grosz and Sidner (1986). The focus tree is a hierarchical representation of all instances of the focus of attention that are expected to appear in users's commands for a given scenario. Sub-focus relations between focus instances are encapsulated in the focus tree: each node, except the root node, represents a sub-focus of its parent node. In addition, we introduced the rules for mapping of the user's commands onto the focus tree, and the algorithms for the transition of the focus of attention. While in the previous chapter we illustrated this approach for the case of the Tangram puzzle, here we apply it for the Tower of Hanoi puzzle. On the semantic level, a command in this puzzle must contain at least two components: the ring that should be moved, and the peg on which the selected ring is to be moved. Each of these components may carry the focus of attention. Therefore, we refer to them as to focus instances. As already discussed, the inspection of the NIMITEK corpus showed that the subjects often produced elliptical commands, i.e., they did not always explicitly utter both focus instances. This may be explained by the fact that focus instances are interrelated. The fragment of the interaction between the subject and the simulated system from the NIMITEK corpus, given in Figure 4.2, illustrates this point.

User$_1$:	*Now the smallest ring . . .*
System$_2$:	selects the smallest disk
User$_3$:	*On the two.*
System$_4$:	puts the selected disk on the second peg

Figure 4.2: Dialogue fragment that illustrates recursive development of the focus of attention.

In the first command (User$_1$), the subject selects the smallest ring; in the second command (User$_3$), she moves the selected ring on the second peg. Considering the focus of attention, in the first command the subject places the focus of attention on the smallest ring. Thereafter she assumes that the selected ring is a part of the shared knowledge between the system and her. Thus, in the second command she introduces a sub-focus of attention that relates to the second peg.

The focus tree that contains all focus instances for the 3-disks version of the Tower of Hanoi puzzle is given in Figure 4.3. The most general focus instance that relates to the puzzle itself is represented by the root node. Nodes on the next level represent focus instances that relate to the rings. Terminal nodes represent focus instances that relate to the pegs. At any given point of the interaction, the focus of attention is placed on exactly one node in the focus tree. We say that this node represents the current focus of attention.

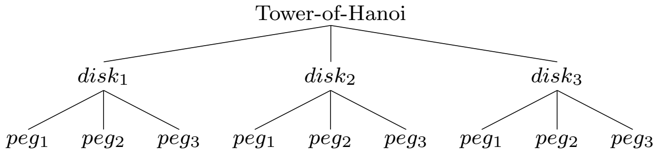


Figure 4.3: The focus tree for the 3-disks version of the Tower of Hanoi puzzle.

4.4.3 The User's Command

In the Wizard-of-Oz experiment described in Chapter 2, no predefined grammar rules for the construction of utterances were given to the subjects. Instead, they were allowed to spontaneously and flexibly formulate their utterances. The analysis of the NIMITEK corpus showed that users' commands may take different syntactic forms (e.g., elliptical or minor commands, context dependent commands, etc.). In the previous chapter we introduced an approach to processing of user's commands for the model of commands contained in the NIMITEK corpus.

Here, for the purpose of defining a dialogue strategy, we introduce another classification of user's commands. When a command uttered by the user is processed, it is assigned to one of the following classes:

- valid command (i.e., the instructed move is allowed according to the rules of the puzzle),
- illegal command (i.e., the instructed move violates the rules of the puzzle, for example, placing of a bigger disk on top of a smaller disk, etc.),
- semantically incorrect command (e.g., the user instructs a non-existing move, for example, trying to move a peg instead of a disk, etc.),

- help command (i.e., the user explicitly asks for support),
- switching between interface languages (German or English),
- unrecognized command (e.g., the user's command is not recognized by the speech recognition module due to background noise, or because it falls outside of the application's domain, etc.).

4.4.4 The State of the User

In our approach to defining a dialogue strategy, we differentiate between three emotional states of the user: *negative*, *neutral* and *positive*. The state of the user is supposed to be detected by the emotion classifier that combines three knowledge resources: prosody (cf. Vlasenko et al. 2007), facial expressions (cf. Niese et al. 2007), and linguistic information of the user's input (cf. Gnjatović et al. 2008e). In this thesis, we do not consider the recognition of emotions in more detail; we assume that the information about the state of the user is delivered to the dialogue management module.

Later in this chapter, we introduce a dialogue strategy that is aimed—among other things—to address the negative emotional state of the user. However, to design and implement an appropriate dialogue strategy, it is necessary to define what the non-neutral user states exactly represent in the given scenario. More precisely, these states should be explained in light of the purpose for which the prototype system was planned in the first place. Therefore, we resort to the NIMITEK corpus in order to get a better insight in possible emotions and emotion-related states of the users. There are two main reasons for this decision. First, the application's domain planned for the prototype system was also used in the WOZ simulation conducted to collect the corpus. Second, the WOZ simulation was especially designed to induce reactions to diverse problems that might occur in the interaction.

The evaluation of the emotional content of the NIMITEK corpus was performed in two phases. The first phase of the evaluation process was reported in Chapter 2 (Section 2.6). It had the primary aim to assess the level of ecological validity of the NIMITEK corpus. This phase demonstrated a satisfying level of ecological validity of the corpus. The subjects signaled genuine emotions overtly and there was a diversity of signaled emotions, emotion-related states and talking styles, as well as a diversity of their intensities. The results of this evaluation phase served as a point of departure for the second evaluation phase whose aim was to define a data-driven model of user states for the given scenario. The important fact of the first evaluation phase is that the choice of annotation labels was data-driven—the evaluators were allowed to introduce labels according to their own perception. Thus, some of the introduced labels represent different but closely related emotions or emotion-related states. We mention a few of these relations between the labels, according to the explanations given by the evaluators:

- The labels *confused* and *insecure* are closely related to the label *fear* graded with low intensity of expressed emotion. This relation is even more obvious if we keep in mind that the label *fear* was never graded with high intensity during the evaluation process.
- The label *disappointed* is closely related to the label *sadness* graded with low intensity of expressed emotion.
- The label *pleased* is closely related to the labels *joy* (graded with low intensity) and *surprised*.

Therefore, there was a need to group labels that relate to similar or mixed emotions or emotion-related states. Following clarifications collected from the evaluators, we mapped these labels onto six classes that form the ARISEN model of user states, as shown in Figure 4.1.

Table 4.1: The ARISEN model of user states.

Class	***Mapped labels***
Annoyed	anger, nervousness, stressed, impatient
Retiring	fear, insecure, confused
Indisposed	sadness, disappointed, accepting, boredom
Satisfied	joy, contentment, pleased
Engaged	thinking, surprised, interested
Neutral	neutral

To prove the appropriateness of this mapping, we performed the second phase of the evaluation. The experimental sessions evaluated in the first phase were re-evaluated in the second phase by a new group of six student evaluators. All evaluators were native German speakers and naïve, i.e., without educational background that relates to the evaluation process (e.g., psychology, linguistics, sociolinguistics, etc.). There are two main differences in the second evaluation phase with respect to the process of evaluation:

- The set of annotation labels was predefined. The evaluators could use only labels from the ARISEN model.

- The re-evaluation was performed over smaller evaluation units. In the first phase, the evaluation unit was a dialogue turn or a group of several successive dialogue turns. The evaluation material was divided in 424 evaluation units. Such units, that are rather long in duration, were selected to demonstrate that emotional expressions are extended in time. In the second evaluation phase, we used finer selection of units—the same evaluation material was divided in 2720 evaluation units.

Each evaluation unit was evaluated by four or five evaluators. They performed the perception test independently of each other. To each evaluation unit evaluators assigned one or more labels from the ARISEN model. Similar as in the first evaluation phase, we used majority voting in order to attribute labels to evaluation units. If at least three evaluators agreed upon a label, it was attributed to the evaluation unit. The evaluation results are given in Table 4.2.

Table 4.2: Results of the second evaluation phase.

Evaluation units	***Number***
with no majority voting	315 (11.58%)
with one assigned label	1907 (70.11%)
with two assigned label	476 (17.5%)
with three assigned label	22 (0.81%)
total	2720 (100%)

Label	***Eval. units attributed with the label***
Annoyed	487 (17.9%)
Retiring	111 (4.08%)
Indisposed	156 (5.74%)
Satisfied	106 (3.9%)
Engaged	1548 (56.91%)
Neutral	517 (19.01%)

As mentioned above, for the purpose of defining a dialogue strategy we differentiate between two non-neutral user states: negative and positive. Using the ARISEN model, we can now define these states as follows:

- Negative state—The user is frustrated due to problems that occurred in the interaction, discouraged because she does not know how to solve a given task, or there is a lack of interest in the user's attitude to solve the task. This includes the user states *Annoyed*, *Retiring* and *Indisposed.*

- Positive state—The user is motivated to solve the task and/or satisfied with the development of interaction. This includes the user states *Engaged* and *Satisfied.*

We comment this briefly. A distinction between the user states based on the valence of the signaled emotion is obviously an important one. The system should be capable to recognize negative user states as indicators for diverse problems that may occur in the interaction. On the other hand, discussing the arousal, it should be kept in mind that it is not expected that a system such as the NIMITEK prototype system would normally provoke emotional reactions of high intensity (at least, if it is not deliberately planned). It is much more likely that signalled emotions would relate to everyday emotions that are inherently less intensive (e.g., *nervousness*, *pleased*, *insecure*, etc.). Therefore, we decided to use the simple categorical labeling of emotion-related content, instead of dimensional one that includes dimensions of valence and arousal.

Now we can answer the question from the beginning of this subsection. The convincing fact that the most frequently marked state from the ARISEN model is *Engagement* points out that the level of engagement of the user towards a given task should be considered as important. A dialogue strategy designed to support the user of the NIMITEK prototype system should address the negative user state on two tracks: (i) to help a frustrated user to overcome problems that occur in the interaction, and (ii) to motivate a discouraged or apathetic user.

4.4.5 The History of Interaction

The history of interaction is a linear data structure that collects relevant information related to the interaction from its beginning. Every time when a new event in the interaction arises (e.g., user's command is performed, some kind of support is provided to the user, etc.), a new entry is added in the history of interaction. An entry comprises the following information:

- current values of other interaction features, i.e., the state of the task, the user's command, the focus of attention, and the state of the user;

- description of the currently applied dialogue strategy;

- time of making the entry.

Collected information is used, among other purposes, to process context dependent user's commands (e.g., "undo", "move the next disk", etc.), to assess the progress of the state of the task towards the final state (e.g., to detect moment when the state of the task draws back from the expected final state, etc.), as well as to dynamically adapt a dialogue strategy, as described below.

4.5 An Adaptive Dialogue Strategy

In this section we introduce an adaptive dialogue strategy implemented in the NIMITEK prototype system to support users while they solve tasks in a graphics system. As mentioned above, the aim of this dialogue strategy is to help users to overcome problems that occur in the interaction and to address the negative user state. The main idea can be formulated as:

> The dialogue manager should dynamically adapt its dialogue strategy according to the actual state of the interaction.

We introduce three requirements that underlie dynamical adaptation of a dialogue strategy. The first requirement is that the user should be provided with useful and sufficient information. The problems that occur in the interaction may be various. They may relate to a given task itself (e.g., the user does not understand the rules of the puzzle, the user does not know how to solve the puzzle, etc.) or to the interface language (e.g., the user does not know how to formulate a valid command, etc.). Information provided to the user should be tailored to a particular problem. Support should be informative enough to help the user to overcome the problem, appropriately emphasized in order to avoid information overload, and clearly presented.

The second requirement is that support should be timely provided. Discussing the annotation of subjects' dialogue acts from the NIMITEK corpus in Chapter 3 (Section 3.3), we noticed that questions make only 4.29% of all utterances produced by the subjects (cf. Table 3.2). In addition, the subjects explicitly demanded support from the system in only 12 of 6798 commands (cf. Table 3.3), although the human operator playing the role of the system offered support 59 times explicitly using the word *help*, e.g., *Do you need help?* (in German: *Brauchen Sie Hilfe?*). Thus, a dialogue strategy should not rely on the assumption that the user will clearly state a need for support. The system should rather detect such a need and be initiator and carrier of provided support.

Finally, the third requirement is that the manner of providing support should be tailored to meet the user's needs. Problems that occur in the interaction may frustrate, indispose or discourage the user. The manner of providing support should be in accordance with the emotional state of the user. To illustrate, let us assume that the user instructed a move that draws back the state of the task from the expected final state. For the user in neutral emotional state, just a warning given by the system might be enough. But, for the user in negative emotional state, the system should probably propose the next correct move in order to prevent further regression of the state of the task and a potential consequential deepening of the negative user state. Nevertheless, the system should not be over-supportive. Providing support in a bad moment may irritate the user or cause negative effects with respect to the user's learning processes. If the user is deeply engaged in solving a given problem or simply exploring interface possibilities of a system, she should not be interrupted by the system, although she may be trying moves that are not optimal or even not correct. Therefore, the user in positive state—e.g., the motivated user or the user that is satisfied with current interaction experience—should be left alone in attempt to find a possible solution for a given problem. And when a negative change of emotional state is detected during this process, support may be provided.

These requirements can be considered to be of general nature. Here, we illustrate them for the Tower of Hanoi puzzle. Design and implementation of the adaptive dialogue strategy in the NIMITEK prototype system includes three distinct but interrelating decision making processes:

- When to provide support to the user?
- What kind of support to provide?
- How to provide support?

The following subsections consider these decision making processes in more detail.

4.5.1 Decision 1: When to Provide Support to the User?

Following Batliner et al. (2000), the user behavior is supposed to mirror the state of the interaction: in the case of troubles in the interaction, the user's behavior changes accordingly. We extend this assumption. The system provides support in four general cases:

- ***A problem related to the task is detected.*** Such problems are divided in two classes:

  - *The user does not understand the rules of the puzzle.* This subcase is recognized when the user utters an illegal command or a semantically incorrect command.
  - *The user cannot solve the puzzle.* This subcase is recognized when the history of interaction shows that the state of the task either draws back from the expected final state or does not make any significant progress towards the final state.

- ***A problem related to the interface language is detected.*** This is related, for example, to the case when user's instruction cannot be recognized.

- ***There is an external trigger for support.*** For example, the user explicitly asks for support.

- ***A negative change of the emotional state of the user is detected.***

In addition, it should be noted that sets of cues that signal different classes of problems are not necessarily disjoint. This brings us to the next decision making process.

4.5.2 Decision 2: What Kind of Support to Provide?

We consider three kinds of support that can be provided to the user:

- *Support related to the task* ***(Task-Support)***, provided when the user does not know how to solve the puzzle or when she violates the rules of the puzzle.

- *Support related to the interface language* ***(Interface-Support)***, provided when the user knows what to instruct but does not know how to formulate it, or when she instructs a non-existing move.

- *Support related to the user's state* ***(User-Support)***, provided when the user is in negative emotional state.

To differentiate between these kinds of support, six conditions are considered according to the diagram given in Figure 4.4.

Values of the first three conditions are determined by the interaction feature *user's command* (cf. Subsection 4.4.3). In case of an illegal command (condition 2), Task-Support is provided. In the case of an unrecognized

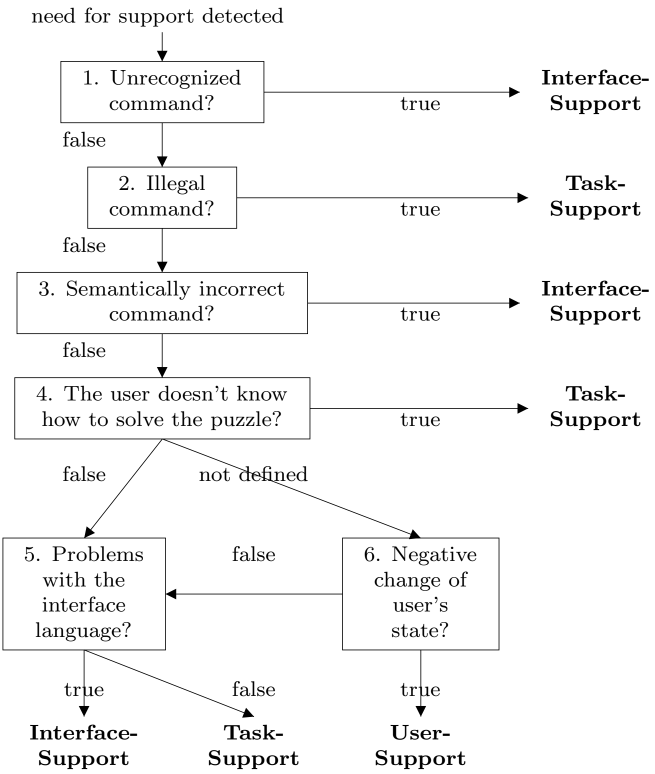


Figure 4.4: Determining the kind of support.

command (condition 1) or a semantically incorrect command (condition 3), Interface-Support is provided.

The case when the user does not know how to solve the puzzle (condition 4) is determined by the interaction feature *history of interaction*. This condition has three possible outputs:

- *true:* when the history of interaction shows that the state of the task either draws back from the expected final state or does not make any significant progress towards the final state. Then, Task-Support is provided.
- *false:* when the history of interaction shows that the state of the task advances towards the solution. Then, condition 5 is considered.
- *not defined:* when it cannot be differentiated between previous two cases based on the information available in the history of interaction

(for example, there is no enough information in the history of interaction). Then, condition 6 is considered.

The case when the user has problems related to the interface language (condition 5) is determined by the interaction feature *focus of attention.* When the focus of attention is placed on an inner node of the focus tree, it signals that the user has started to formulate a command, but still has not completed the formulation. In this case, Interface-Support is provided. Otherwise, when the focus of attention is placed on the root node or a terminal node of the focus tree, it may signal that the user did not start to formulate a command or that she finished the formulation of a command. In this case, Task-Support is provided.

The condition related to the negative user's state (condition 6) is determined by the interaction feature *state of the user.* If the user is in negative emotional state, User-Support is provided. Otherwise, the interaction feature focus of attention (condition 5) is considered, as discussed above.

4.5.3 Decision 3: How to Provide Support?

When the kind of support is determined, the system decides in what manner it should be provided. The manner is determined by the state of the user. We differentiate between three cases:

- *No support* for users in positive emotional state.
- *Low intensity of support* for users in neutral emotional state.
- *High intensity of support* for users in negative emotional state.

Low intensity of Task-Support means to inform the user that her last move pushed her away from the final solution of the puzzle or that her last move violates the rules of the game. High intensity of Task-Support is to inform the user as well, but also to propose the next move. This move is determined by the state of the task.

The aim of Interface-Support is to help the user to complete a command. Providing low intensity of this support, the system guides the user to complete the started command by stating iterative questions (e.g., which disk should be selected, where to move the selected disk, etc.). High intensity of Interface-Support is to check whether the started command can be completed in such a way that it pushes the state of the task towards the final solution. If so, the system proposes such command to the user. Otherwise, the system warns the user that the started command is not appropriate.

User-Support is provided only to the user in negative emotional state. As this support is not connected with problems related to the puzzle or to the interface language, its main purpose is to address the negative emotional state of the user. When the user in negative emotional state makes a correct move, i.e., the move that advances the state of the task towards expected solution, the system occasionally produces short, encouraging messages. In addition, the user's emotional state is always mirrored by displaying animation of one of three *emoticons* that represent facial expressions of positive, neutral and negative emotional states, respectively.

It should be kept in mind that sometimes the emotion recognition module does not provide any information about emotional state of the user. In such cases, the unknown user's state is treated as being neutral.

4.6 Dialogue Management Module

In this section, we provide a brief overview of the functionality of the dialogue management module and of its relations to functionalities of other modules incorporated in the NIMITEK prototype system. Adaptive dialogue management in this system encapsulates functionalities of three sub-modules:

- natural language understanding module,
- attentional state module,
- dialogue strategy module.

These three sub-modules implement theoretical considerations introduced in this thesis.

Processing and performing of the user's command is represented in Figure 4.5. The textual version of the user's command outputted from the speech recognition module is delivered to the emotional classifier and to the natural language (NL) understanding module. The emotion classifier is supposed to combine three knowledge resources: prosody, facial expression, and linguistic information. Thus, input to the emotions classifier consists of the audio stream, the video stream, and the textual version of the user's input. Detected emotional state of the user is outputted from the emotion classifier.

The natural language understanding module extracts focus instances from the textual input, interprets the command, as discussed in Chapter 3, and forwards it (shown by the dashed arrows):

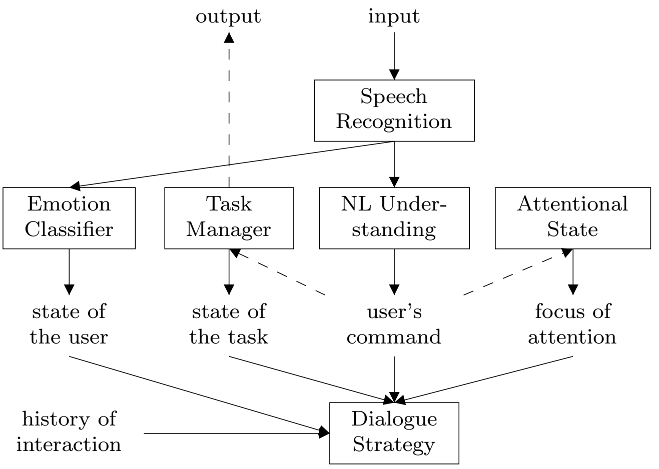


Figure 4.5: Processing of the user's command.

- to the task manager module (including the graphical platform) for a performance of the interpreted command, for an update of the state of the task, and for an appropriate graphical display,
- to the attentional state module for an update of the focus of attention.

In addition, a new entry is added to the history of interaction, containing: updated state of the task, detected command, current focus of attention, detected state of the user, and current time.

Then, if support should be provided, the system applies the dialogue strategy according to the current state of the interaction, as described above. Providing support to the user is represented in Figure 4.6. Generally, support information may contain a proposed move and an audio message accompanied by textual output of the message content on the screen. In the case when support contains only an audio message, this information is delivered to the task manager module for an appropriate display. If support contains also a proposed move, this information is send:

- to the task manager module for a performance of the proposed command and an update of the state of the task,
- to the attentional state module for an update of the focus of attention.

In addition, the history of interaction is updated with the following information: kind and intensity of support, content of support (e.g., proposed

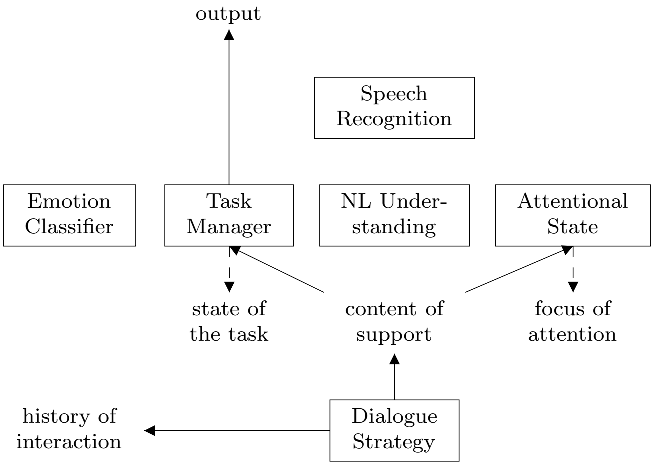


Figure 4.6: Providing support.

command, etc.), updated state of the task, updated focus of attention, and current time.

4.7 Conclusion

This chapter proposed an approach to designing adaptive dialogue strategies. More precisely, this chapter reported about design and implementation of the adaptive dialogue strategy in the NIMITEK prototype spoken dialogue system for supporting users while they solve a problem in a graphics system. However, this approach is not limited to the interaction domain of the NIMITEK prototype system only. It covers the class of spoken dialogue systems that are intended to manage a subclass of task-oriented dialogues, i.e., dialogues that are primarily concentrated on a given task, where the state of the task is observable in the sense that it can be explicitly defined and evaluated with respect to how it corresponds to expected final states.

The underlying idea is that the system dynamically adapts its dialogue strategy according to the actual state of the interaction. For the purpose of this contribution, we introduced the state of the interaction as a composite of five interaction features: the state of the task, the user's command, the focus of attention, the state of the user, and the history of interaction. The appropriate attention was devoted to the discussion about the meaning of these interaction features (especially of the state of the user) in the given application's scenario.

We introduced three requirements that underlie dynamical adaptation of the dialogue strategy. The first requirement is that the user should be provided with useful and sufficient information tailored to a particular problem. Support should be informative enough to help the user to overcome the problem, appropriately emphasized in order to avoid information overload, and clearly presented. The second requirement is that support should be timely provided. It means that a dialogue strategy should not rely on the assumption that the user will clearly state a need for support. The system should rather detect such a need and be initiator and carrier of provided support. Finally, the third requirement is that the manner of providing support should be tailored to meet the user's needs. The manner of providing support should be in accordance with the emotional state of the user. Therefore, dynamical adaptation of the introduced dialogue strategy is determined by three distinct but interrelated decision making processes: When to provide support? What kind of support to provide? How to provide support?

Finally, we provided a brief overview of the functionality of the dialogue management module and of its relations to functionalities of other modules incorporated in the NIMITEK prototype system. Adaptive dialogue management in this system encapsulates functionalities of three sub-modules: natural language understanding module, attentional state module, and dialogue strategy module. These three sub-modules implement theoretical considerations introduced in this thesis.

The next chapter discusses the functionality of the dialogue management module in more detail. Among other illustrations, we give an analysis of an actual dialogue between the user and the NIMITEK prototype system that took place during the testing of the system.

Chapter 5

Discussion

5.1 Introduction

In the previous chapters, we introduced and illustrated important theoretical considerations related to adaptive dialogue management in human-machine interaction, and reported the implementation of the dialogue management module in the NIMITEK prototype spoken dialogue system that exemplifies them. However, several important questions remain to be discussed. Some of them are: How do the proposed algorithms—both for transition of the focus of attention and for applying the adaptive dialogue strategy—work in a realistic scenario? The implemented dialogue strategy provides *immediate* support, e.g., proposing the next correct move—what about long-term support, e.g., helping the user to understand the recursive nature of the puzzle? To what extent is the modeling approach task-independent? Can we extend it to cover more domains simultaneously? How is it to be used in the framework of the introduced adaptive dialogue strategy? How may the model of attentional state and the adaptive dialogue strategy contribute to overcome some limitations of automatic speech recognition technology?

This chapter addresses these questions. First, in order to illustrate important points of the implementation of the model of attentional state and of the adaptive dialogue strategy, we analyze an actual dialogue between the user and the NIMITEK prototype system that took place during the testing of the system. Second, we introduce and discuss the implementation of an extension of the adaptive dialogue strategy aimed to provide long-term support to the user. Finally, we discuss and illustrate how the dialogue management module handles miscommunication on different levels: the conversational level, the intentional level, and the signal level.

5.2 An Example

The theoretical concepts that underlie the implementation of the dialogue management module in the NIMITEK prototype system (i.e., the model of attentional state and the adaptive dialogue strategy) were thoroughly elaborated in two previous chapters. Here we do not repeat this discussion. The aim of this section is to illustrate the functionality of the dialogue management module in more detail. We give an analysis of an actual dialogue between the user and the NIMITEK prototype system that took place during the testing of the system. The whole dialogue is given in Figures 5.1, 5.5 and 5.6. Utterances produced by the user and the system are written in *italic*, descriptions of non-verbal actions performed by the system (e.g., moving a disc, etc.) are underlined, and moments when the system detects a change of the state of the user are given in **bold**. In the following subsections, we discuss processing of different types of user's commands, providing support to the user and the multilingual working mode.

5.2.1 Processing of User's Commands

In Chapter 3 we introduced an approach to processing of user's commands in human-machine interaction for the restricted model of commands contained in the NIMITEK corpus. We illustrated it for the Tangram puzzle (cf. Section 3.5) and, in addition, discussed the issues of phrasal lexicon and generalizability (cf. Section 3.6). We stated that the implementation of the proposed model of attentional state within the dialogue management module in the NIMITEK prototype system is independent of changes of the structure of the focus tree (e.g., a change from the Tangram puzzle to the Tower of Hanoi puzzle, etc.) and of changes of the vocabulary. In this subsection we provide some illustrations for this statement. We consider the implementation of the introduced model of attentional state for another domain—the Tower of Hanoi puzzle—and discuss processing of different types of user's commands. The discussion on generalizability is continued below in Section 5.3.2.

User's commands may take different syntactic forms. Referring to the first part of the dialogue, shown in Figure 5.1, we discuss processing of *elliptical*, *verbose* and *context dependent* user's commands. Processing of commands are closely related to the notion of the focus of attention introduced in Chapter 3, and crucially supported by the focus tree structure (cf. focus tree for the Tower of Hanoi puzzle given in Figure 4.3 in the previous chapter). Transitions of the focus of attention that are relevant for the observed dialogue fragment are represented in Figures 5.2, 5.3 and 5.4. Changes of a temporary focus of attention are marked with dashed arrows. Nodes representing temporary focus of attention are positioned in ovals, while nodes representing the focus of attention after processing of the user's command are positioned in boxes.

System$_1$:	**classifies the user state as neutral**
User$_2$:	*Den kleinsten Ring ...* (engl.: *The smallest ring ...*)
System$_3$:	selects the smallest disk
User$_4$:	*... auf die Drei.* (engl.: *... on the three.*)
System$_5$:	puts the selected disk on the third peg
User$_6$:	*Ich möchte gerne die mittlere Scheibe auf Position Zwei schieben.* (engl.: *I would like to move the middle disk to position two.*)
System$_7$:	moves the middle disk to the second peg
User$_8$:	*Den nächsten Ring ...* (engl.: *The next ring ...*)
System$_9$:	selects the biggest disk

Figure 5.1: Dialogue between the user and the NIMITEK prototype system: Part 1.

(1) Processing elliptical commands. The commands User$_2$ (*The smallest ring ...*) and User$_4$ (*... on the three*) are elliptical—in each of them the user gives only a part of the information that is expected to be contained in a fully formulated command. In command User$_2$, the system recognizes a focus instance that relates to the smallest disk. Then it checks if this focus instance can be represented by some of the descendant nodes of the node representing the current focus of attention. For the given starting focus of attention (placed on the root node of the focus tree at the beginning of the dialogue), there is such a node—$disk_1$. The new focus of attention is placed on this node. The same discussion holds for the command User$_4$. A focus instance that relates to the third peg is recognized. This focus instance can be represented by three nodes in the focus tree: peg_{13}, peg_{23} and peg_{33}. However, only one of them is a descendant node of the node representing the current focus of attention—peg_{13}. Thus, the new focus of attention is placed on this node. The transition of the focus of attention is illustrated in Figure 5.2.

(2) Processing verbose commands. The command User$_6$ (*I would like to move the middle disk to position two*) is verbose in the sense that it contains words that are not a part of the vocabulary recognized by the

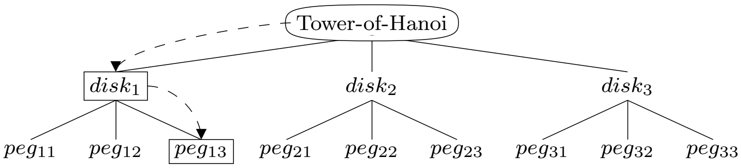


Figure 5.2: Transition of the focus of attention for the commands User$_2$ and User$_4$.

speech recognition module. The textual version of this command outputted from the speech recognizer may be represented as:

<not recognized> the middle disk <not recognized> position two

We make a small digression related to the issue of automatic speech recognition. The set of phrases recognized by the speech recognition module was intentionally restricted to phrases that relate to focus instances. A question that may arise is whether the problem of ”verbose” commands could be solved by improving or extending the speech recognition module. However, the answer is negative for minimum two reasons. The first reason is that it cannot be expected that users will always produce ”well structured” utterances that contain only predefined words and phrases. The second reason relates to state-of-the-art automatic speech recognition technology. Non-accurate speech recognition is still a frequent phenomenon in human-machine interaction. In settings when systems operate under realistic conditions (e.g., spontaneous speech, large vocabularies and user population, etc.) average word recognition error rates are 20–30% for native speakers (Bohus and Rudnicky, 2008, p. 123–4). Moreover, non-accurate speech recognition can cause miscommunication on different levels of interaction. The issue is discussed in Section 5.4.

Now we get back to the processing of commands User$_6$. The system recognizes two focus instances. The first focus instance, *the middle disk*, can be represented by the node $disk_2$; the second focus instance, *position two*, by the nodes peg_{12}, peg_{22} and peg_{32}. However, none of these nodes are a descendant node of the node representing the current focus of attention (i.e., peg_{13}). Thus, a temporary focus of attention is moved, in a bottom-up manner, towards more general focus instances. It is iteratively transited to the closest antecedent node of the node peg_{13} that satisfies the condition that its descendant nodes can represent all focus instances from the command—in this case it is the root node. Then, similarly as already explained above, all

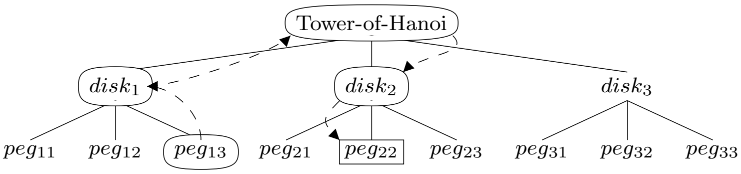


Figure 5.3: Transition of the focus of attention for the command User$_6$.

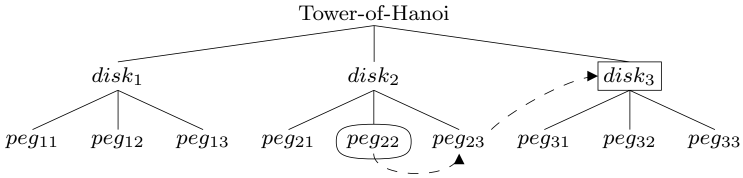


Figure 5.4: Transition of the focus of attention for the command User$_8$.

changes of a temporary focus of attention are directed towards more specific focus instances. The focus of attention is first placed on the node $disk_2$ and then to its child node peg_{22}. The transition of the focus of attention is illustrated in Figure 5.3.

(3) Processing context dependent commands. The command User$_8$ (*The next ring* ...) is elliptical, but also context dependent. Observing this command *in abstracto*, i.e., isolated from the surrounding dialogue context, it can be assumed that a disk should be selected, but it is not specified which actual disk. Therefore, the contextual information should be taken into account. The previously selected disk was the middle disk. Thus, the phrase *the next ring* relates to the biggest disk and the new focus of attention is placed on the node $disk_3$. The transition of the focus of attention is illustrated in Figure 5.4 (cf. explanation of processing of context dependent commands in Section 3.6).

5.2.2 Supporting the User

Let us observe now the continuation of the dialogue between the user and the system, given in Figure 5.5. In the command User$_{10}$ (... *on the three*) the user tries to instruct the system to place the (previously selected) biggest disk to the third peg. However, according to the rules of the puzzle, it is not possible to place a larger disk onto a smaller one. Instruction of an illegal command is an indicator for the system that the user may need support. It should be kept in mind that here the system classifies the emotional state of the user as neutral, so it decides only to inform the user that the instructed move is not possible (System$_{11}$).

In command User$_{12}$ (*Help*), the user explicitly asks for support. This is another indicator for the system. The current focus of attention is still on the node *disk*$_3$ (it has not been changed since the user's command User$_8$). Since the focus of attention is placed on an inner node of the focus tree, it signals that the user has started to formulate a command, but still has not completed the formulation (i.e., if the focus of attention was placed on a terminal node of the focus tree, it would signal that the user has finished the formulation of a command, cf. Subsection 4.5.1). In addition, it assumes that the user wants to formulate such a command that would place the focus of attention on one of the terminal nodes from the sub-tree determined by the node in the current focus of attention as its root node. Therefore, in System$_{13}$, the system asks the user on which peg does he want to place the selected ring. The user refuses to answer the system's question and repeats the request for support in User$_{14}$ (*I said Help!*) It should be noted that now the system classifies the emotional state of the user as negative (System$_{15}$), so its response to the user's demand is not the same as in the previous case. The system decides to provide support of higher intensity—it informs the user that the moving of the selected disk would draw him back from the correct solution of the puzzle and propose the user to select another disk (System$_{16}$).

The user responds on the system demand. In command User$_{17}$ (*The smallest ring ...*), he instructs the system to select the smallest ring. The system performs this instruction (System$_{18}$), and the user then again asks for support (User$_{19}$). The emotional state of the user is still classified as negative, so support of higher intensity is given again. However, in this case, the selected ring was the correct choice. Therefore, in System$_{20}$, the system proposes to place the selected ring on the second peg, and asks the user to confirm that he agrees with the proposal (in System$_{21}$). The user approves the proposal in User$_{22}$. A similar discussion holds for the rest of the observed dialogue fragment. In User$_{24}$, the user instructs a command that draws him back from the correct solution of the puzzle. The system reacts and proposes the user to undo this move (System$_{25}$).

This dialogue fragment demonstrates how the dialogue manager dynamically adapts its dialogue strategy according to the current state of the interaction. As introduced in the previous chapter, the system takes into account interaction features in order to decide when to provide support, which kind of support to provide, and in which manner. In the given fragment, the system decided to provide support to the user in the following cases: the

User$_{10}$: *... auf die Drei.* (engl.: *... on the three.*)

System$_{11}$: *Dieser Schritt ist nicht möglich. Sie dürfen keinen Ring auf einen kleineren legen.* (engl.: *This move is not possible. You cannot place a larger disc onto a smaller one.*)

User$_{12}$: *Hilfe.* (engl.: *Help.*)

System$_{13}$: *Welchen Stab möchten Sie auswählen?* (engl.: *Which peg do you want to select?*)

User$_{14}$: *Ich sagte Hilfe!* (engl.: *I said help!*)

System$_{15}$: **classifies the user state as negative**

System$_{16}$: *Der letzte Schritt führte von der Lösung weg. Ich schlage vor, dass Sie einen anderen Ring auswählen.* (engl.: *The last move drew you back from the solution. I propose that you select another disc.*)

User$_{17}$: *Den kleinsten Ring ...* (engl.: *The smallest ring ...*)

System$_{18}$: selects the smallest disk

User$_{19}$: *Hilfe.* (engl.: *Help.*)

System$_{20}$: *Ich möchte diesen Schritt vorschlagen.* (engl.: *Let me propose the following move.*)
moves the smallest disk to the second peg

System$_{21}$: *Sind Sie einverstanden?* (engl.: *Is this ok?*)

User$_{22}$: *Ja.* (engl.: *Yes.*)

System$_{23}$: **classifies the user state as neutral**

User$_{24}$: *Den kleinsten Ring auf die Eins.* (engl.: *The smallest disk on the one.*)

System$_{25}$: *Ich möchte diesen Schritt vorschlagen.* (engl.: *Let me propose the following move.*)
returns the smallest disk again to the second peg

System$_{26}$: *Sind Sie einverstanden?* (engl.: *Is this ok?*)

User$_{27}$: *Ja.* (engl.: *Yes.*)

Figure 5.5: Dialogue between the user and the NIMITEK prototype system: Part 2.

Table 5.1: Example for dynamical adaptation of the dialogue strategy.

Dialogue turn	***Changes in the state of the interaction***	***System's response***
System_{11}	*user's command:* illegal command	informing the user
System_{13}	*user's command:* help command	helping the user to formulate the command
System_{16}	*state of the user:* negative	warning the user that the intended move is misleading
System_{20-1}	*state of the task:* disk selected *user's command:* help command *focus of attention:* $disk_1$	proposing the next move
System_{25-6}	*state of the task:* disk moved *state of the user:* neutral *focus of attention:* peg_{12}	undoing the user's last move

user instructed a command that violates the rules of the puzzle; the user did not complete the formulation of a command; the user explicitly asked for support; and the user instructed a move that draws him back from the expected final state of the puzzle. Support was provided in five dialogue turns, as summarized in Table 5.1. The table shows which interaction features were changed—with respect to the previous state of the interaction—prior to system's decisions to provide support, and the corresponding responses of the system.

For example, the manner in which the system provides support is determined by the emotional state of the user. In two successive commands User_{12} and User_{14}, the user explicitly asks for support. The states of the interaction differ only with respect to the emotional state of the user—the system classifies the state of the user as neutral for the first command, and as negative for the second command. Common for both requests is that the user previously selected a wrong disk to be moved, but he still did not specify the peg on which the disk should be moved. In the former case, the system decides not to interfere with the user's intention. It just supports the user to finish the formulation of the command that he started. In the latter case, the system tries to prevent the already frustrated user from instructing a wrong move. It warns the user that moving of the selected disk would draw him back from the correct solution of the puzzle, and proposes the user to select another disk.

5.2.3 Multilingual Working Mode

In Chapter 3, we noted that sets of phrases that are assigned to focus instances are defined in input XML files, independently of the implementation of the algorithms for transition of the focus of attention. This means that the implementation of the dialogue management module is independent of changes of the vocabulary, including translation of phrases from German into English. These changes do not require a change in the core implementation, but just a redefinition of input XML files. This is illustrated in the final fragment of the observed dialogue given in Figure 5.6.

Verbal interaction with the NIMITEK prototype system is supported in two languages: German and English. In commands User$_{28}$ (*English*) and User$_{32}$ (*German*), the user instructs the system to switch to English and German language, respectively. Thus, the commands User$_{30}$ and User$_{32}$ were uttered in English.

User$_{28}$:	*Englisch.* (engl.: *English.*)
System$_{29}$:	switches to English language
User$_{30}$:	*The big ring to the column number three.*
System$_{31}$:	moves the biggest ring to the third peg
User$_{32}$:	*German.*
System$_{33}$:	switches to German language
User$_{34}$:	*Den kleinsten Ring auf die Eins.* (engl.: *The smallest disk on the one.*)
System$_{35}$:	moves the smallest disk to the first peg

Figure 5.6: Dialogue between the user and the NIMITEK prototype system: Part 3.

Further advantages of the "implementation-independent" definition of the structure of the focus tree and of sets of phrases that are assigned to focus instances are discussed and illustrated in Subsection 5.3.2 in more detail.

5.3 Long-Term Task-Support

The implemented dialogue strategy provides a kind of support that we can refer to as *immediate* or *short-term*. For example, if the user does not know how to solve the puzzle, the next correct move may be proposed by the system. However, besides providing such immediate support, the system does not try to help the user to understand the recursive nature of the Tower of Hanoi puzzle. In this section, we discuss the issue of addressing the user's attitude towards a given task on the *long-term*.

The aim of long-term Task-Support is to help the user to understand the recursive nature of the Tower of Hanoi puzzle, instead only to provide immediate help. For example, if the user does not know how to solve the puzzle, it may be more appropriate to bring the user to the concept of the puzzle by facing her with a less complex version of the puzzle. And if the user shows a good performance in solving the puzzle, the system may propose a more complex version of the puzzle. In addition, the user should have a possibility to explicitly demand a change of the level of complexity of the puzzle. This section introduces and discusses an extension of the adaptive dialogue strategy aimed to provide long-term Task-Support to the user.

5.3.1 Extension of the Dialogue Strategy

In this subsection, we introduce an extension of the adaptive dialogue strategy introduced in the previous chapter that relates to Task-Support.

This extension may be described as follows: At the beginning of the interaction, the user is faced with the 3-disks version of the Tower of Hanoi puzzle. If she successfully completes the puzzle, and the system provided Task-Support less then three times during this period, the user is congratulated and offered to solve a more complex, 4-disks version of the puzzle. Otherwise, if the puzzle is still not solved and Task-Support should be provided for the third time, the system does not provide support immediately. It proposes first the user to try to solve a less complex version of the puzzle before she continues with the current version of the puzzle. If the user accepts this offer, the system stores the state of the puzzle and starts the 2-disks version of the puzzle. When the user completes this version of the puzzle, the system restores the state of the 3-disks version of the puzzle, so that the user may continue with it. If, however, the user does not accept the offer, she is allowed to continue with the current version of the puzzle and Task-Support with high intensity (cf. Subsection 4.5.3 in the previous chapter) is provided, regardless of the emotional state of the user.

Let us now make a small digression and consider for a moment the case when the user accepts the offer to start a less complex version of the puzzle. One possibility is that the user successfully completes the 2-disks version of the puzzle and then returns to the previous version of the puzzle. Another possibility is that the user, in the process of solving the 2-disks version of the puzzle, concludes that she understood the recursive nature of the puzzle and that there is no need to solve it up to the end, so she interrupts it and demands to return immediately to the previous version of the puzzle.

However, in both cases—whether the user accepted or refused the system's offer to start a less complex version of the puzzle—the user will at some moment continue with the 3-disks version of the puzzle. The behavior of the system when it detects a need for Task-Support next time (i.e., the fourth time in the scope of the 3-disks version of the puzzle) is:

- If the user completed the 2-disks version of the puzzle, the system does not offer her to go back. Instead, from that point of interaction until the puzzle is solved, it provides Task-Support—when a need for this kind of support is detected—always with high intensity, regardless of the emotional state of the user.

- If the user interrupted the less complex version of the puzzle or even did not accept the system's offer to start it, the system repeats its offer. If the user again refuses the offer or interrupts the less complex version of the puzzle, the system will not repeat the offer in later situations when a need for Task-Support may be detected. Similarly as in the former point, the system will provide Task-Support with high intensity, regardless of the emotional state of the user.

It should be noted that the user has at any time the possibility to demand a change to the less complex or to the more complex version of the puzzle.

The system's behavior when the user experiences task related problems while she solves the 2-disks and the 4-disks versions of the puzzle remains to be described. The dialogue strategy is similar, but slightly modified. For the 2-disks version, a further simplification of the puzzle is not specified. Therefore, from the moment when a need for Task-Support is detected for the third time during the solving of the given version of the puzzle, the system starts to provide it with high intensity, regardless of the emotional state of the user. For the 4-disks version, a degradation to the 3-disks version is reasonable only if the user skipped to the 4-disks version without previously having finished the 3-disks version. If this is the case, the system behaves in the same manner as for the 3-disks version of the puzzle. Otherwise, it behaves in the same manner as for the 2-disks version of the puzzle.

5.3.2 Again on Modeling Attentional Information

This section continues the discussion on the issue of generalizability of the model of attentional state introduced in Chapter 3. First, we make a small recapitulation. In Chapter 3, we proposed an approach to processing of user's commands in human-machine interaction for the restricted model of commands contained in the NIMITEK corpus. We introduced the concept of the focus tree in order to model attentional information on the level of a user's command and the rules for transition of the focus of attention for different types of user's commands. In Section 3.6 we discussed that the implementation of the proposed model of attentional state within the dialogue management module in the NIMITEK prototype system is independent of changes of the structure of the focus tree and of changes of the vocabulary. To support this discussion, processing of commands was illustrated for two dialogue domains: the Tangram puzzle (Section 3.5) and the Tower of Hanoi puzzle (Subsection 5.2.1).

This section provides another illustration to support this discussion. We consider processing of the user's commands in the context of the introduced extension of the dialogue strategy. In the scope of the interaction, the user may be engaged in solving three versions of the Tower of Hanoi puzzle. Points of departure related to processing of users' commands can be summarized as:

- We have three different domains of interaction, where attentional information from each of them can be modeled by a focus tree.
- During the interaction between the user and the system, the current focus of attention may be moved from one domain to another. For example, the user can switch between different versions of the puzzle.
- The sets of phrases assigned to focus instances from different domains are not necessarily disjunctive. For example: The phrase ”smallest disk” could refer to the smallest disks in all three versions of the puzzle. Furthermore, the phrase ”biggest disk” may refer to the second disk in the 2-disks version of the puzzle, to the third disk in the 3-disks version of the puzzle, or to the fourth disk in the 4-disks version of the puzzle.

Let F_2, F_3 and F_4 be focus trees for the 2-disks version, for the 3-disks version, and for the 4-disks version of the puzzle, respectively. We model attentional information for all these three domains by a more general focus tree F that encompasses F_2, F_3 and F_4. This focus tree is illustrated in Figure 5.7. Its root node does not represent a focus instance from any of the domains. This ”abstract” root node serves to encompass F_2, F_3 and F_4 as sub-trees in F: root nodes of the focus sub-trees F_2, F_3 and F_4 are positioned as child nodes of the root node of F.

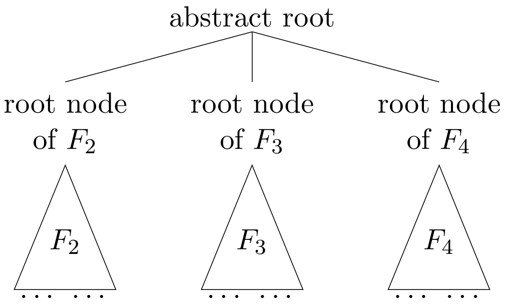


Figure 5.7: Illustration of a focus tree that encompasses focus instances from several domains of interaction.

It is important to note that rules for transition of the focus of attention introduced in Chapter 3 (cf. algorithms introduced in Section 3.5) hold also for the case of such an "encompassing" focus tree. When we extended the domain of interaction from the scenario with only one version of the Tower of Hanoi puzzle to the scenario with tree versions of this puzzle, we changed the structure of the focus tree (as shown in Figure 5.7) and extended the phrasal lexicon (i.e., sets of phrases related to focus instances). We recall again that the structure of the focus tree and sets of phrases that are assigned to focus instances are defined in input XML files independently of the implementation of the dialogue management module. Therefore, seen from the aspect of processing the user's commands, no changes in the core implementation of the proposed model of attentional state within the dialogue management module were necessary. We just had to change the input XML files that define (i) the structure of the focus tree and (ii) the sets of phrases assigned to focus instances in the focus tree.

For the purpose of completeness, we note that two minor changes of the existing implementation of the dialogue management module had been made. These changes were necessary due to the definition of the dialogue strategy. They are:

- According to the dialogue strategy, the interaction starts with the 3-disks version of the Tower of Hanoi puzzle. Therefore, at the start of the interaction, the current focus on attention is placed on the root node of F_3.

- According to the dialogue strategy, the user may choose to stop her activity in one version of the puzzle, spend some time in solving another version of the puzzle, and than later come back to the initial version. Therefore, when the user changes between different versions of the puzzle, the systems stores the state of the puzzle, so that it can be restored later when the user continues with the initial version. However, changing between different versions of the puzzle implies a transition of the focus of attention from one sub-tree to another, e.g., starting the 2-disks version of the puzzle places the focus of attention on the root node of the sub-tree F_2. It is thus necessary to store also the current focus of attention, so that it can be later restored in order to support processing of the user's commands after she comes back to the initial version of the puzzle.

5.4 Miscommunication

In the context of spoken natural language human-machine interaction, miscommunication can occur on different levels: on the conversational level (e.g., the user's utterance falls outside of the system's functionality), on the intentional level (e.g., the user's utterance falls outside of the system's semantic grammar), on the signal level (e.g., inaccurate speech recognition), etc. Miscommunication is a frequent and natural phenomenon in spoken communication and appears to be unavoidable (McTear et al., 2005). It is clear that the state-of-the-art automatic speech recognition (ASR) approaches still cannot deal with flexible, unrestricted users' language. Also, it is not reasonable to expect that users will always behave "cooperatively" and produce utterances that fall within the application's domain, scope and grammar. Forcing users to always produce "well structured" utterances would significantly limit the naturalness of the interaction. Furthermore, for users in affected states, such a cooperative behavior is hardly to be expected at all.

Still, left unmediated by better error awareness and recovery mechanisms, miscommunication may *severely limit the naturalness of the interaction and the complexity of the tasks that can be addressed* (Bohus and Rudnicky, 2008, p. 124). Therefore, in order to achieve a habitable language interface, there is an essential need for both models of interaction and dialogue strategies that support the user to overcome problems that occur due to miscommunication. This section gives an overview of related works in the field of handling miscommunication in spoken dialogue systems, and then discusses properties of the model of attentional state and of the adaptive dialogue strategy introduced in this thesis that can reduce the level of miscommunication.

5.4.1 Background and Related Work

Observing the current state-of-the-art in ASR technology, Lee (2007) concludes that the research community has yet to address a number of challenges. He states some limitations (Lee, 2007, p. 25):

- *Restrictive systems:* to effectively utilize spoken language applications, the users have to follow a strict set of protocols.

- *Fragile technology:* careful designs have to be rigorously practiced to hide technology deficiencies.

- *Low accuracy:* ASR accuracies often degrade dramatically in adverse conditions to an extent that applications become unusable even for cooperative users. ASR usually gives much larger error rates than human speech recognition (HSR)—in highly noisy conditions even more than one order of magnitude higher than HSR.

Bohus and Rudnicky (2008, p. 123–4) state that in settings when systems operate under the conditions of spontaneous speech, large vocabularies and user population, and large variability in input line quality, average word recognition error rates are 20–30%, and they go up to 50% for non-native speakers. They note that speech recognition errors can cause two types of understanding errors in a spoken dialogue system:

- misunderstandings—when the system obtains an incorrect interpretation of the user's input,

- non-understandings—when the system fails to obtain any interpretation of the input.

Most work on miscommunication in the context of spoken natural language human-machine interaction has been focussed on miscommunication caused by inaccurate recognition (McTear, 2008, p. 101). This is in line with results of the empirical analysis that 62% of non-understandings and 77% of misunderstandings originate at the speech recognition level (Bohus and Rudnicky, 2008, p. 131–2). However, miscommunication is not only caused by word recognition errors. Bohus and Rudnicky (2008, p. 131) identify four main sources of errors:

- On the conversation level—The user's utterance falls outside of the application's domain (e.g., the user asks the room-reservation system about the weather) or outside of the application's scope (e.g., the user asks whether a room has windows).

- On the intention level—The user's utterance falls outside of the system's semantic grammar (e.g., the user utters ”erase reservation”, which is not in the system's grammar, instead of ”cancel reservation”, which is in the system's grammar).

- On the signal level—The user's utterance is misrecognized or not recognized by the ASR module of the system, although it is within the application's domain, scope and grammar.

- On the channel level—The end-pointer is not able to correctly segment the incoming audio signal (e.g., the microphone truncates the user's utterance).

Bohus and Rudnicky (2008, p. 125–8) compared the individual performance of various non-understanding recovery strategies in the domain of spoken dialogue system that handles conference room reservations. This research was designed as *a between-group experiment* with two conditions. Participants in the first condition interacted with a version of the system that used a random policy to engage recovery strategies. Participants in the second condition interacted with a modified Wizard-of-Oz version of the same system where the human operator decided which strategy to apply. The available recovery strategies in both conditions were: asking the user to repeat the utterance, asking the user to rephrase the utterance, repeating the previous prompt, notifying the user that a non-understanding has occurred, advancing the task by moving on to a different question, telling the user what she can say at that point of dialogue, providing a longer help message explaining the current state of the dialogue, etc. Bohus and Rudnicky (2008, p. 151) report that the best performing dialogue strategies in the observed domain were:

- Advancing the conversation by ignoring the non-understanding and trying an alternative dialogue plan.

- Providing help messages containing sample responses for the current system question.

McTear et al. (2005, p. 249–250) starts with the assumption that errors are a natural occurrence in spoken communication and thus unavoidable.

They argue that an approach to handling miscommunication is required to have methods for detecting and dealing with miscommunication when it occurs. Their approach are based on *the theory of grounding, which states that participants in a conversation collaborate to establish and maintain common ground and thus seek to avoid miscommunication and to deal with it appropriately when it occurs.* McTear et al. state that:

> [. . .] error handling involves deciding what to do when an error is detected (or suspected). In order to make such decisions, the system makes use of all the information it has available to it and assesses the costs and benefits of repairing the miscommunication. This information, often referred to as the system's information state [. . .], may include information about what has been said in the dialogue so far, current agendas and priorities, recognition confidence levels, and various other sources of information that together contribute to the agent's dialogue strategy (McTear et al., 2005, p. 250).

This approach has been implemented in the Queen's Communicator, the system that handles transactions in the domains of accommodation requests and booking, and event requests and booking (O'Neill et al. 2003, McTear et al. 2005, McTear 2008). McTear (2008, p. 115) gives an illustration: The system's decision of which confirmation strategy (e.g., implicit, explicit) to use is based on the state of information to be confirmed (e.g., new for the system, inferred by the system, repeated by the user, modified by the user, negated by the user, etc.) and its degree of confirmedness that varies according to whether the user had just repeated the information, negotiated it, modified it, etc. Taking this information into account, the system uses a set of rules to determine its confirmation strategy.

A similar approach is proposed by Skantze (2008a,b). He reports a discourse modeler for conversational spoken language, called GALATEA, designed *to support concept-level error handling.* It *tracks the grounding status of concepts that are mentioned during the discourse, i.e. information about who said what when* (Skantze, 2008b, p. 156) and builds a discourse model that may be consulted by the action manager to select an error handling strategy.

5.4.2 Miscommunication on the Conversational and the Intentional Level

In Chapter 3, we discussed that the user's commands may take different syntactic forms, and proposed an approach to processing of such commands. In this chapter (Subsection 5.2.1), we illustrated processing of elliptical commands, verbose commands, and context dependent commands. Here, we summarize some advantages of this approach to processing of commands that can reduce the level of miscommunication on the conversation level and on the intentional level:

- The user can utter diverse phrases in order to address the same entity. For example, the focus instance *smallest disk* can be referred to as "smallest disk", "first disk", etc.

- The user can change the order of phrases in utterance. Given sets of phrases that may be used (defined in an input XML file, cf. Section 3.6 and Subsection 5.3.2), the system automatically derives focus instances from the user's command by detecting phrases that relate to certain focus instances. The order of phrases in utterance is not important. For example, the utterances "the smallest disk on the third peg" and "on the three—the smallest disk" are interpreted by the system in the same way.

- The user can use *wrapper* expressions (cf. Subsection 3.3.1) that fall outside of the system's semantic grammar. In contrast to focus instances, wrappers do not relate to propositional content of the user's commands. For example, wrappers that represent expressions of politeness in the following user's utterances are given in *italic*: "The middle disk *please* on the number two" and "*I would like* to put the smallest disk on the three". To interpret propositional content of the user's commands, the system derives only phrases that relate to focus instances, while it ignores wrappers. Finally, it should be mentioned that wrappers may carry affect information, so they are important for recognition and tracking of the user's emotional state from linguistic information, as illustrated and discussed in Appendix A.

5.4.3 Miscommunication on the Signal Level

On the signal level, miscommunication is caused by inaccurate speech recognition (i.e., non-recognition and misrecognition). In the previous subsection, we state that non-recognition of wrappers in the user's command do not affect interpretation of propositional content. However, the same does not hold for non-recognition or misrecognition of parts of the user's command that relate to focus instances, because they carry information about the propositional content.

This subsection discusses how the implemented dialogue strategy handles miscommunication. It is important to note that the implemented dialogue strategy encapsulates two conceptual ideas, mentioned also in Subsection 5.4.1, that relate to miscommunication handling:

- The conversation should be advanced in spite of miscommunication. The system should support the user to overcome problems that occur due to miscommunication.
- Support should be dynamically adapted according to the current state of the interaction.

In order to make these statements more clear, let us assume that a valid user command is not correctly recognized by the automatic speech recognition module. Due to actual error in speech recognition, the dialogue management module may interpret the command as a valid—although misinterpreted—command, an illegal command, a semantically incorrect command, or an unrecognized command. These classes of commands were introduced in Subsection 4.4.3. For each of these classes, the system provides—if needed—support (cf. Figure 4.4 in Chapter 4). E.g., in case of an illegal command, Task-Support is provided; in case of an unrecognized command or a semantically incorrect command, Interface-Support is provided, etc.

For the purpose of illustration and without loss of generality, let us also assume that the user uttered "the second disk on the first peg". We discuss the system's behavior in several cases of inaccurate speech recognition:

- *Case 1: A part of the command is correctly recognized, and at least one phrase that relates to a focus instance can be derived, while the rest of the command is not recognized.* For example, the textual version of the command outputted from the speech recognizer may be "the second disk <not recognized>". This command will be processed as elliptical (cf. Subsection 5.2.1), but still valid command. The second disk will be selected, and the user will have a possibility to specify a peg again in the next command. Another example for the textual version of this command outputted from the speech recognizer is "<not recognized> on the first peg". In this case, depending on the state of the task (i.e., which disk is currently selected), the interpreted move (i.e., moving the selected disk on the first peg) may be valid and thus performed, or illegal (i.e., not according to the rules of the puzzle). In the former case, if the performed move was not the move that the user instructed, she has a possibility to instruct an undo command. Also, if the performed move pushed the state of the task away from the final solution of the puzzle, Task-Support is provided by the system, as described in the previous chapter. In the latter case, if the move was interpreted as illegal, appropriate Task-Support is again provided by the system.

- *Case 2: The command is not recognized, or a part of the command is correctly recognized, but no phrase that relates to a focus instance can be derived.* For example, the textual version of the command outputted from the speech recognizer may be "<not recognized> disk <not recognized> peg". According to the definition of the dialogue strategy, when an unrecognized command is detected, Interface-Support (introduced in Chapter 4, Section 4.5) is provided. The aim of Interface-Support is to help the user to formulate a command, e.g., the system may guide the user to formulate a command by stating iterative questions: which disk should be selected, where to move the selected disk, etc.

- *Case 3: A part of the command or the whole command is misrecognized.* An example of the textual version of the command outputted from the speech recognizer may be "the second disk on the first disk". This will be classified as a semantically incorrect command. Another example is "the second disk on the second peg" which—depending of the current state of the interaction—may be classified as a valid command or an illegal command. Discussion of these examples is similar as above. If the command is classified as a semantically incorrect command or as an illegal command, the appropriate support is provided. In the case when the command is classified as valid, it is performed, although the user did not instruct that particular command. However, even then, if the performed move drew the state of the task back from the final solution of the puzzle, support will be provided (cf. Section 4.5). Only in the case when a misrecognized command is interpreted as a valid command that pushes the state of the task towards the final solution, the misrecognition will not be detected by the system. However, this is not a critical oversight, since it advances the conversation in a good direction.

In order to illustrate how the implemented dialogue strategy handles miscommunication, these cases were purposely selected so that they and their combinations cover a wide range of different interaction situations caused by inaccurate speech recognition.

5.5 Conclusion

This chapter discussed and illustrated various aspects of functionality of the adaptive dialogue management module in the NIMITEK prototype spoken dialogue system. In the first part of the chapter, we analyzed an actual dialogue between the user and the prototype system that took place during the testing of the system. We illustrated important points of the implementation: processing of the user's commands of different syntactic forms (e.g., elliptical commands, verbose commands, context dependent commands), adaptive dialogue strategy that supports the user to overcome various problems that occur in the interaction, and multilingual working mode.

The second part introduced and discussed the implementation of an extension of the adaptive dialogue strategy aimed to provide long-term Task-Support to the user. The system varies the level of complexity of the puzzle to help the user to understand the recursive nature of the puzzle. Built on the top of the existing implementation of the dialogue strategy (introduced in Chapter 4), this extension is also designed to be adaptive. Furthermore, we considered processing of the user's commands in the context of the introduced extension of the dialogue strategy. We introduced a generalization of the focus tree structure and discussed that it can be used to process the user's commands in the case of more domains of interaction that may be concurrently involved in a given dialogue instance. It is important to note that rules for transition of the focus of attention introduced in Chapter 3 hold also for this case, and that—seen from the aspect of processing the user's commands—no changes in the existing implementation of the dialogue management module were necessary. Redefinition of the structure of the focus tree and extension of sets of phrases that relate to different focus instances were achieved by simple redefinition of input XML files.

Finally, the third part of the chapter discussed how the dialogue management module handles miscommunication on the conversational level (e.g., the user's utterance falls outside of the system's functionality), on the intentional level (e.g., the user's utterance falls outside of the system's semantic grammar), and on the signal level (e.g., inaccurate speech recognition). We pointed out and illustrated advantages of the model of attentional state introduced in Chapter 3 that can reduce the degree of miscommunication on the conversation level and on the intentional level, e.g., the user can use diverse phrases in order to address the same entity, the user can change the order of phrases in utterance, etc. In addition, we discussed how the implemented dialogue strategy handles miscommunication on the signal level. The underlying idea is that the system tries to advance the interaction in spite of miscommunication by providing support—adapted according to the current state of the interaction—to the user. For example, due to actual error in speech recognition, the dialogue management module may interpret a valid user's command as a valid—although misinterpreted—command, an illegal command, a semantically incorrect command, or an unrecognized command. For each of these classes of commands, the system provides—if needed—support according to the definition of the dialogue strategy (cf. Chapter 4, e.g., in the case of an illegal command, Task-Support is provided; in the case of an unrecognized command or a semantically incorrect command, Interface-Support is provided, etc.). The system's response is illustrated by several examples that were selected so that they and their combinations cover a wide range of different interaction situations caused by inaccurate speech recognition.

Chapter 6

Conclusion

This thesis introduced an approach to adaptive dialogue management in human-machine interaction. It has made contributions in the areas of theory, experimental practice, and system development.

We discussed important theoretical considerations and implementation issues in development of an adaptive dialogue management module, and exemplified them for the NIMITEK prototype spoken dialogue system. At the application level, the dialogue management module in the NIMITEK prototype system was designed and implemented to support users while they solve a task in a graphics system (e.g., the Tower of Hanoi puzzle). More generally, this module illustrates the focal points of adaptive dialogue management presented in this thesis, e.g., interpreting propositional content of the users commands, modeling contextual information, dynamically adapting the dialogue strategy, providing support, etc.

It is important to note that the proposed approach to adaptive dialogue management is not limited to the interaction domain of the NIMITEK prototype system only. With respect to the domain of the interaction, this approach covers the class of spoken dialogue systems that are intended to manage a subclass of task-oriented dialogues, i.e., dialogues that are primarily concentrated on a given task, where the state of the task is observable in the sense that it can be explicitly defined and evaluated regarding to how it corresponds to expected final states. In addition, we concentrated on spoken human-machine interaction in the specific case where some kind of display with a graphical interface is involved. We discussed that display represents an additional non-linguistic context shared between the user and the system, and that it may influence the language of the user (e.g., predominant use of elliptical and minor utterances, context dependent utterances, etc.). With respect to the processing of the user's spoken input of different syntactic forms, the proposed approach covers the class of spoken dialogue systems that are intended to control a subclass of graphical user interfaces, e.g., manipulating with graphical entities represented on the display, controlling graphical menus, solving graphically-based tasks, playing interactive board games that includes spatial reasoning, etc.

The introduced approach to adaptive dialogue management represents an integration of several lines of research: producing and evaluating corpora of affected behavior in human-machine interaction, modeling attentional information on the level of the user's command, and designing adaptive dialogue strategies.

(1) Producing ecologically valid emotional corpora. It is a widely accepted fact that research on the role of emotions in human-machine interaction is essentially supported by corpora containing samples of emotional expressions. One of the fundamental requirements for such corpora is that they have to be ecologically valid, i.e., collected samples should be representative of emotions as they occur in everyday life. The main criticism of existing corpora is leveled against the often used practice of using material produced by actors and disregarding less intense, everyday emotions.

We addressed these methodological desiderata. Chapter 2 addressed research questions of producing and evaluating an emotional corpus, and presented the NIMITEK corpus[1] of affected behavior in human-machine interaction collected within the reported research. During the process of collecting the corpus, proper attention was devoted to the issue of its ecological validity. We proposed two additional requirements that are to be met in order that a WOZ scenario designed to elicit affected behavior could result in ecologically valid data. First, subjects have to be motivated to accomplish a given task in order that a successful accomplishment or a failure to accomplish could induce an emotional state. This requirement is introduced to address the problem of non-engaged subjects that are role-playing only. The second requirement for the successful emotion elicitation in WOZ experiments is that subjects have to be stimulated to express their emotions overtly. In addition, we discussed a need for a more sophisticated approach to dialogue management, concluding that experimental settings should allow experimenters to observe and control the development of the dialogue between subjects and the simulated system. Implications of this observation on wizard's dialogue strategies and response time were considered. Also,

[1] Please note that the NIMITEK corpus is available from the authors for research purposes upon request.

possible pitfalls of the proposed requirements were discussed.

The NIMITEK corpus contains 15 hours of audio and video recordings produced during a refined Wizard-of-Oz experiment designed to induce emotional reactions. The evaluation of the corpus with respect to its emotional content demonstrated a satisfying level of ecological validity. The corpus contains recordings of genuine emotions that were overtly signaled. It is not oriented to extreme representations of a few emotions only but comprises also expressions of less intense emotions. Emotional expressions of diverse emotions are extended in modality (voice and facial gesture) and time. Different classes of non-neutral talking style are marked in the obtained data. In addition to audio and video recordings of the experimental sessions, all dialogues in the corpus are transcribed, and dialogue acts are annotated.

(2) Naturalness of the interface language. One of the widely accepted postulates of human-machine interaction is that it should be as natural as possible. An important aspect of naturalness of the interaction is certainly naturalness of the interface language. The essence of naturalness of the interface language is that users can express themselves without conscious effort to follow rules of a predefined grammar while producing their utterances. Inspection of the NIMITEK corpus showed that the subjects often produced irregular (e.g., elliptical or minor, etc.) commands. Forcing users to always produce "regular" utterances would be too restrictive and not well accepted. It can not be expected that users—and especially users in affected states—will always behave cooperatively and produce utterances that fall within the application's domain, scope and grammar. This implies that a language interface should be able to cope with various dialogue phenomena related to the users' language, such as different syntactic forms of users' utterances (from syntactically very simple utterances to verbose utterances), high frequency of ungrammaticalities, use of ellipses, context dependent utterances, etc.

This issue is addressed in Chapter 3. Attentional information is already recognized as crucial for processing of utterances in discourse. We motivated and introduced a new model of attentional state—the focus tree. We used it to model attentional information on the level of the user's command and to introduced rules for transition of the focus of attention. Three main advantages were gained from this modeling approach. They are discussed and illustrated in Chapter 3 as well as in Chapter 5. Here, we summarize them. First, instead of predefining a grammar for accepted users' commands, we allow flexible formulation of commands. The implementation was demonstrated to work well for different syntactic forms of commands: elliptical commands, verbose commands (that are only partially recognized by the speech recognition module), and context dependent commands. Second, the proposed modeling method and algorithms are not *a priori* related to some specific predefined task. The introduced algorithms for transition of the focus of attention are independent of the structure of the focus tree and of the content of the phrasal lexicon. That makes the implementation of the proposed model of attentional state within the dialogue management module in the NIMITEK prototype system independent of (i) changes of the structure of the focus tree (e.g., a change from the Tangram puzzle to the Tower of Hanoi puzzle, etc.), and (ii) changes of the vocabulary (e.g., changing the size of the vocabulary by extending or redefining sets of phrases, changing the language of the vocabulary—processing of users' commands was demonstrated to function for German and English—etc.). Third, our approach to modeling attentional information is not limited only to verbally uttered commands. It supports also non-verbal dialogue acts produced by the user (e.g., using a mouse or a keyboard, etc.) or by the system (e.g., performing a move, etc.).

(3) Adaptive dialogue strategies. The issue of naturalness of human-machine interaction considers more than just the language interface. In addition to it, we stated two requirements that we find to be essential in achieving a higher level of naturalness of the interaction. First, the behavior of the system should be dynamically adapted according to the current state of the interaction. Second, the user should be considered as an integral part of the interaction. Consequently, providing a response, systems should also take properties of the user—especially the emotional state of the user—into account.

Chapter 4 proposed an approach to designing adaptive dialogue strategies, and exemplified it for the adaptive dialogue strategy in the NIMITEK prototype spoken dialogue system for supporting users while they solve a problem in a graphics system. The main idea is that the system dynamically adapts its dialogue strategy according to the current state of the interaction. We defined the state of the interaction as a composite of five interaction features: the state of the task, the user's command, the focus of attention, the state of the user, and the history of interaction. We introduced three requirements that underlie dynamical adaptation of a dialogue strategy aimed to support the user. The first requirement is that the user should be provided with useful and sufficient information, tailored to a particular problem. Support should be informative enough to allow the user to overcome the problem, appropriately emphasized in order to avoid information overload, and clearly presented. The second requirement is that support should be timely provided. Thus, a dialogue strategy should not rely on the assumption that the user will clearly state a need for support. The system should rather detect such a need and be initiator and carrier of provided support. The third requirement is that the manner of providing support should be tailored to meet the user's needs, i.e., it should be in accordance with the emotional state of the user. To briefly illustrate: For the user in neutral emotional state that instructs a "wrong" move, just a warning given by the system might be enough. But, for the user in negative emotional state, the system should probably also propose the next correct move in order to prevent further regression of the state of the task and a potential consequential deepening of the negative user state. Nevertheless, the system should not be over-supportive. Providing support in a bad moment may irritate the user or cause negative effects with respect to the user's learning processes. If the user is deeply engaged in solving a given problem or simply exploring interface possibilities of the system, she should not be interrupted by the system, although she may be trying moves that are not optimal or even not correct.

Design and implementation of the adaptive dialogue strategy in the NIMITEK prototype system includes three distinct but interrelating decision making processes that reflect above requirements: When to provide support to the user? What kind of support to provide? How to provide support? Support is provided in different interaction situation: when the user does not understand the given task, when the user does not know how to solve the given task, when the user's instruction cannot be recognized, when the user explicitly asks for support, etc. The system provides three kinds of support: Task-Support (related to the task itself, e.g., explaining the rules of the puzzle and helping to find its solution), Interface-Support (related to the interface language, e.g., helping to formulate a valid command), and User-Support (addressing negative emotional states of the user, e.g., producing short, encouraging messages). The manner of providing support is determined by the state of the user. No support is provided to the user in positive emotional state (e.g., the user that is deeply engaged in solving a given problem). Support with low intensity is provided to the user in neutral emotional state. Finally, support with high intensity (i.e., more informative support) is provided to the user in negative emotional state.

Finally, Chapter 5 discussed various aspects of functionality of the adaptive dialogue management module in the NIMITEK prototype system. First, we analyzed an actual dialogue between the user and the prototype system, and illustrated important points of the implementation: processing of the user's commands of different syntactic forms, adaptive dialogue strategy that supports the user, and multilingual working mode. Second, we introduced and discussed the implementation of an extension of the adaptive dialogue strategy aimed to provide long-term Task-Support to the user. The underlying idea is that the system varies the level of complexity of the puzzle—by switching between three different versions of the puzzle—to help the user to understand the recursive nature of the puzzle. Furthermore, we considered processing of the user's commands in the context of the extension of the dialogue strategy. We showed how the model of attentional state may be generalized to support processing of the user's commands in the case of more domains of interaction that may be concurrently involved in a given dialogue instance. Third, we discussed how the dialogue management module handles miscommunication on different levels. We pointed out advantages of the model of attentional state that can reduce the degree of miscommunication on the conversation level and on the intentional level. In addition, we discussed how the implemented dialogue strategy handles miscommunication on the signal level (i.e., miscommunication caused by inaccurate speech recognition).

Contributions that have been made by this thesis represent a basis for further investigating both theoretical considerations and implementation issues in the field of adaptive dialogue management in human-machine interaction. Future prospects of research in this field includes several research problems. We shortly state some of them, although the list is by no means complete:

- Investigation of linguistic cues for early recognition of negative dialogue developments.

- Further development of dialogue strategies for preventing and handling negative dialogue development.

- Investigation of the role of empathy in human-machine interaction and of dialogue strategies and linguistic means to convey it.

- Enabling the dialogue manager to use reinforcement learning—e.g., by analyzing the history of interaction and the profile of the user—in order to dynamically adapt its dialogue strategy for a given user in a given situation.

- Investigation of more complex user models and alternative models of emotions.

We take these to be of great importance for increasing the level of adaptivity of human-machine interfaces. Adaptivity is currently one of the main foci of interdisciplinary research efforts to make a *breakthrough* towards *companion-enabled* cognitive technical systems. One of the aims of this thesis is to make a step in this direction.

Appendix A

Linguistic Expression of Emotion

A.1 Introduction

In the following, we employ the NIMITEK corpus (introduced in Chapter 2) as a tool that provides an empirical foundation for analyzing emotional content of linguistic features in the transcribed conversations. We shortly discuss various linguistic features that may carry affect information (e.g., key words and phrases, lexical cohesive agencies, dialogue act sequences, etc.). All examples are translated into English, and German original is given in parentheses.

We start our discussion showing a sequence of commands taken from the NIMITEK corpus, given in Figure A.1. The discourse was produced by the subject while she was solving a graphical puzzle. The human operator that

Subject: *I take the parallelogram ... Yes, move slowly to the right ... More ... stop ... Please move slowly up ... stop ... Please move slowly to the right.*
(*Ich nehme das Parallelogramm ... Ja, langsam nach rechts schieben ... Weiter ... Stopp ... Bitte langsam nach oben schieben ... Stopp ... Bitte langsam nach rechts drehen.*)

Figure A.1: A sequence of commands produced by the subject solving a graphical task.

plays the role of the system in the WOZ settings performs the instructed commands properly, so the interaction between the subject and "the system" unfolds without problems. In contrast to such an "unproblematic" dialogue fragment, we recognized that there are different styles how subjects approach the system in a case when a problem occurs. One could be termed *pedagogical* or *teacherese* and is characterized by trying to teach the computer how it should behave properly. The dialogue fragment given in Figure A.2 illustrates this.

Another style is characterized by open signals of despair and helplessness when problems pile up, as shown in Figure A.3.

We investigate a typology of users' utterances and sequences of users' utterances that signal emotional state of the user. In the next sections, we discuss insights from the NIMITEK corpus relating to different linguistic features that may carry affect information.

A.2 Specific Key Words and Phrases

One way to recognize an emotional state is to detect key words and phrases in users' utterances. Figure A.4 provides some examples of key words and phrases that relate to certain emotion-related states and attitudes.

However, expressions of emotions are not limited to a single dialogue act, but they map over a range of mutually related dialogue acts. Therefore, we consider also lexical cohesive agencies (that relate dialogue acts in a structure, cf. Halliday 1994, p. 309–334) and dialogue acts sequences (cf. Batliner et al. 2000) in order to detect signals of emotion-related states. In the next sections, we are primarily focused on recognizing signals of negative emotional states.

A.3 Lexical Cohesive Agencies

Ellipsis-substitutions. Ellipsis-substitution is a form of anaphoric cohesion in a discourse, *where we presuppose something by means of what is left out* (Halliday, 1994, p. 316). For example, in Figure A.5 the subject replaces the verb *move* (*fahren*) with the general verb *do* (*machen*).

It is important to note that *in ellipsis-substitutions the typical meaning is not one of co-reference. There is always some significant difference between the second instance and the first* (Halliday, 1994, p. 322). To illustrate this, let us observe a typical example for an ellipsis-substitution in the NIMITEK corpus: *Please do it! (Bitte tu das!)*. This utterance does not explicitly carry

Subject: *The smallest ring from the one to the three ... Stop! Back ... Where should the smallest ring go? ... On the three, ... three! ... What are you doing there?*
(*Den kleinsten Ring von der Eins auf die Drei ... Stopp! Zurück ... Der kleinste Ring soll wohin? ... Auf die Drei ... Drei! ... Was tust du da?*)

Wizard: *I am doing just what you are asking.*
(*Ich tue nur, was Sie sagen.*)

Subject: *Incorrect! ... Down ... these rings stay down! ... Understood?*
(*Falsch! ... Runter ... Diese Ringe bleiben liegen! ... Verstanden?*)

Wizard: *What rings?*
(*Welche Ringe?*)

Subject: *The middle ring and the large ring, ... now put ...*
(*Der mittelgroße Ring und der große Ring, ... Jetzt de ...*)

Wizard: *I don't understand you.*
(*Ich versteh' Sie nicht.*)

Subject: *The smallest ring! ... On the three please, ... on the three ... No! On the three ... Where is the three? ... On the three! ... Where should the smallest ring be placed?*
(*Den kleinsten Ring! ... Auf die Drei, bitte, ... auf die Drei ... Nein, auf die Drei ... Wo ist die Drei? ... Auf die Drei! ... Wo soll der kleinste Ring hin?*)

Wizard: *On the three.*
(*Aud die Drei.*)

Subject: *Please, do it!*
Bitte, tu' das!

Figure A.2: A dialogue fragment illustrating *pedagogical* or *teacherese* talk style.

Subject: *No! No! No! . . . Execution not correct!*
(*Nein! Nein! Nein! . . . Ausführung nicht korrekt!*)
Wizard: *You still haven't solved the task.*
(*Sie haben die Aufgabe noch nicht gelöst.*)
Subject: *I know that.*
(*Das weiß ich.*)
Wizard: *Do you need help?*
(*Brauchen Sie Hilfe?*)
Subject: *The communication is not working. You don't understand my instructions and you don't do what I say.*
(*Die Kommunikation stimmt nicht. Du verstehst meine Anweisungen nicht und machst nicht das, was ich sage.*)
Wizard: *I am doing only what you are saying.*
(*Ich mache nur, was Sie sagen.*)
Subject: *No, it is not true.*
(*Nein, das stimmt nicht.*)

Figure A.3: A dialogue fragment illustrating open signals of helplessness.

Annoyed: *Sh*t (Sche*ße), stupid (blöd), Do what I say (Tu was ich sage), I've had enough of it (Es reicht mir), It is mean (Das ist gemein). Oh . . . something like this I hate just like the plague. (Ooohh . . . so was hasse ich doch wie die Pest.)*
Retiring: *I don't understand it (Ich versteh' das nicht), It's not working at all (Das geht doch gar nicht), I don't understand the task (Ich versteh' die Aufgabe nicht).*
Indisposed: *I am going now (ich geh' gleich), Oh man (Oh man), God (Gott), I don't feel like doing any more. (Ich hab' kein' Bock mehr.)*
Offending: *You think, doll. (Denkst du, Puppe.)*
Satisfied: *Super (Super), yeah! (yeah!), awesome (geil), I am good, am I not? (Bin gut, was?)*
Polite: *Please (Bitte), I would like (Ich hätt' gern).*
Friendly: *Dear computer, . . . (Lieber Computer, . . .)*

Figure A.4: Examples of key words and phrases that relate to various emotional states.

Subject: *Why are you* ***moving*** *it on peg 2? Why? Why are you* ***doing*** *this step?*
(*Warum* ***fährst*** *du auf Säule 2? Warum? Warum* ***machst*** *du diesen Schritt?)*)

Figure A.5: An example of a dialogue sequence containing question with ellipsis-substitution.

information what is the system expected to do. It contains an elliptical-substitution (*do*), a reference (*it*) that relates to context and an element of politeness (*please*). Here the ellipsis-substitution is used to signal that the action that the system performed is not the same as the action instructed by the user. Thus, ellipsis-substitutions may carry a signal of a potential problem in the interaction. Such a problem may be related to the given task or to the interface language.

Lexical cohesion. The choice of lexical items to create cohesion in the discourse can also signal an emotion-related state, both on the lexical level (e.g., repetitions), as well as on the semantic level (e.g., reformulations). This is illustrated in Figure A.6.

Simple repetition:	*It just cannot be. It just ... It just cannot be. (Das kann doch nicht sein. Das ist doch ... das kann doch nicht sein.)* *What is the problem? What is the problem? (Was ist das Problem? Was ist das Problem?)*
Repetiton and remark:	*Left up. Left up. Left up.* ***I said*** *left up. (Links oben. Links oben. Links oben.* ***Ich habe gesagt*** *links oben.)*
Reformulation:	**Not true** at all. That's **definitely wrong**. (Gar **nicht wahr**. Das **stimmt gar nicht**.)

Figure A.6: Examples illustrating how the choice of lexical items to create lexical cohesion can relate to negative user states.

Dialogue act sequences. The type of dialogue acts in a sequence may also carry affect information. For example, a sequence of questions may signal potential problems in interaction. An illustration of such a sequence containing a question with ellipsis substitution is given in Figure A.5.

A.4 Questions

Subjects in the NIMITEK corpus used also questions to signal their frustration or uncertainty. Common for these questions is a relatively high level of abstraction, as we illustrate below. We differentiate several groups of such questions. The first group of questions signals the frustration of the user. It contains probe questions that start with 'what' or 'why', usually characterized by short structure, and containing ellipsis substitutions, e.g.:

> *What are you doing? (Was tust du?) What's the point of that? (Was soll das?) Why are you doing this? (Warum machst du das?)*

Questions from the second group relate to a concrete action that was or should be performed by the system, e.g.:

> *Why don't you move the 8 to left? (Warum schiebst du die 8 nicht nach links?)*

The third group signals that the user is confused or retiring. These questions usually contain a reference to a previous utterance, such as:

> *But there is one more, isn't there? (Aber es gibt noch eins, oder?)*

Finally, the fourth group contains rhetorical questions, e.g.:

> *Did I say disk downward? No! (Hab' ich gesagt Scheibe nach unten? Nein!)*

A.5 Negation

Negation, combined with other functional elements (e.g., modal particles) may also signal a potential problem in interaction. Figure A.7 provide some examples.

A.6 Conclusion

Recognition of emotion from linguistic information should not be limited to analysis of specific key words and phrases. Approaches that are based only on detection of "emotional" keywords and phrases are related to various problems (cf. Wu et al. 2006). We illustrate some of them using examples from the NIMITEK corpus.

Negation:	*Right up. **No**, right up. (Rechts oben. **Nein**, rechts oben.)*
Negation with enhancement:	***No**, it is not right, It is **simply** not right. (**Nein**, das stimmt nicht. Das stimmt **einfach** nicht.)* *I don't understand it. I **really** don't understand it. (Ich verstehs nicht. Ich verstehs **echt** nicht.)*
Negation with confirmation:	*Do I mean it? **No**, I don't mean it, **do I**? (Meinte ich das? **Nee**, das meinte ich nicht, **oder**?)*
Negation with generalization:	***No**, that will **all** come to nothing. **No**, that will come to nothing. (**Nee**, das wird **alles** nichts mehr. **Nee**, das wird nichts mehr.)*

Figure A.7: Examples illustrating how negation can relate to negative user states.

Subject: Yes ... It is given this way ... System, what is the solution of this task? ... Repeat the task. (Ja ... Das steht da auch so drin ... System, wie ist die Lösung dieser Aufgabe? ... Aufgabe wiederholen.)

Figure A.8: A sequence of commands that carry *negative* prosody.

The lack of prosodic information. This is illustrated by the sequence of the subject's utterances given in Figure A.8. Due to its prosodic cues, this sequence was attributed by the human evaluators with the labels *Annoyed* and *Retiring.* However, the transcript of this sequence does not contain any obvious "emotional" key word that indicates affect information.

Ambiguity in defining emotional keywords and phrases. For example, the subject's exclamation "Oh" may express both surprise and disappointment in the NIMITEK corpus.

Ambiguity in syntactic and semantic information. For example, in the following sequence of the subject's utterances:

> Downward ... System, repeat the instructions! (Nach unten ... System, wiederhole die Anweisungen!)

the system should resolve what the meaning of the second utterance is. It may be that the subject does not understand the given task, so she asks the system to repeat introducing instructions. Or it may be that a problem related to the interface language occurred, so the user, as a part of her dialogue strategy, asks the system to repeat the instructions previously uttered by the subject in order to control if the system understood them correctly. In the latter case, it is also a signal of a potential problem in interaction.

Therefore, additional linguistic features, such as these illustrated above (e.g., information about structure of dialogue acts, context, lexical information, etc.), should also be considered in the recognition process. In addition, an automatic annotator for recognition and tracking of the user's emotional state from linguistic information and other linguistic features should be combined with other classifiers (e.g., prosodic classifier, facial expression classifier, etc.).

Appendix B

The Test Given in the WOZ Experiment

B.1 Introduction

The test given to the subjects in the Wizard-of-Oz experiment described in Chapter 2 consists of 14 graphically-based tasks that can be classified in 6 groups: Filling empty place (3 tasks), Classification (2 tasks), the Tangram puzzle (3 tasks), the Grid puzzle (3 tasks), the Tower of Hanoi puzzle (2 tasks), the Three Jugs Problem (1 task). The specification of these tasks is given below.

B.2 Filling Empty Place

- **Task description:** The subject should select one of the four pictures given on the right side of the screen that logically fits into the field marked with the question mark.

- **Introduction made by the wizard:** *Bitte wählen Sie eines der vier Teile auf der rechten Seite. Sagen Sie dann, ob es in das Feld mit dem Fragezeichen passt.*

- Three versions of the task are given in Figures B.1, B.2 and B.3.

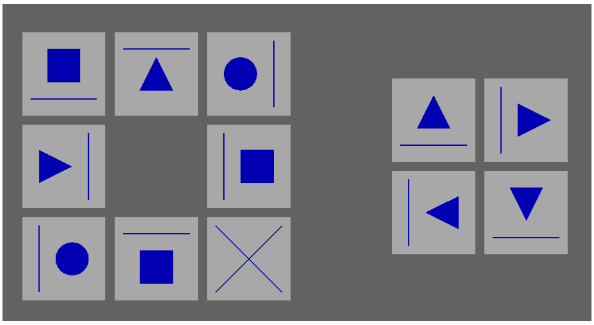


Figure B.1: The first version of the task ”Filling empty place”.

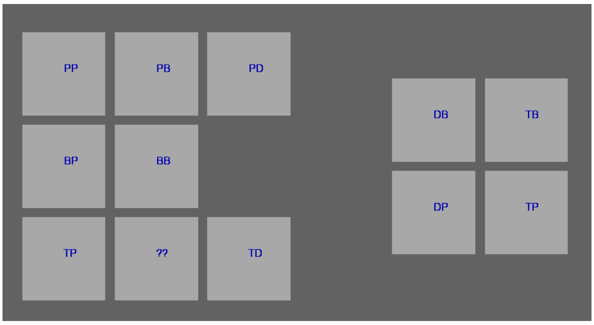


Figure B.2: The second version of the task ”Filling empty place”.

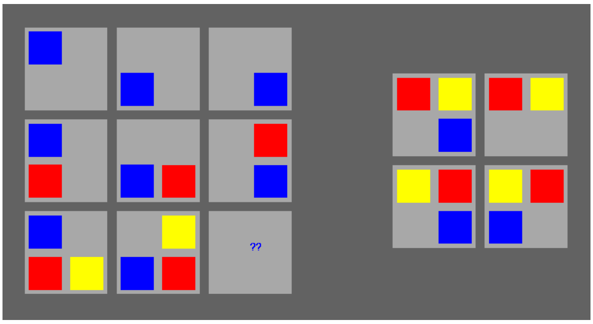


Figure B.3: The third version of the task ”Filling empty place”.

B.3 Classification

- **Task description:** A 3D-figure and a group of 2D-nets are presented to the subject. For each 2D-net the subject should say whether it represents the given 3D-figure unfolded in 2D or not.

- **Introduction made by the wizard:** *Auf der linken Seite befindet sich ein dreidimensionaler Körper. Auf der rechten Seite sind mögliche Lösungen angegeben, wie die Oberfläche des Körpers entfaltet werden könnte. Jede Lösung kann entweder falsch oder richtig sein. Bitte sagen Sie für jede der sechs möglichen Lösungen, ob sie eine korrekte Entfaltung des dreidimensionalen Körpers ist.*

- Two versions of the task are given in Figures B.4 and B.5.

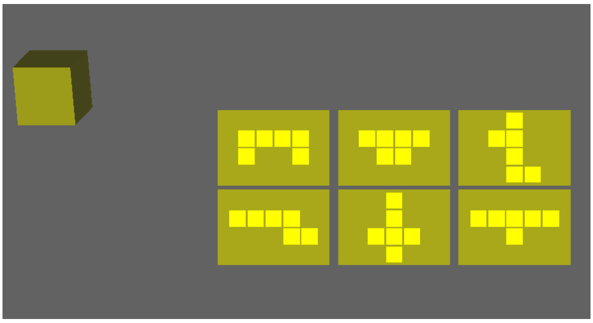


Figure B.4: The first version of the task "Classification".

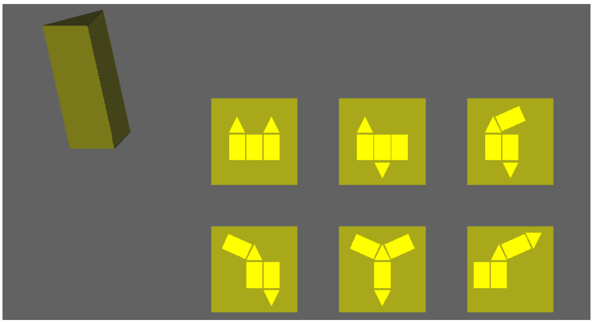


Figure B.5: The second version of the task "Classification".

B.4 Tangram Puzzle

- **Task description:** The Tangram puzzle consists of 7 pieces: five triangles, a square and a parallelogram. The objective of this puzzle is to form a given shape using all 7 pieces. The pieces must not overlap.

- **Introduction made by the wizard:** *Auf der linken Seite sehen Sie die sieben Teile eines Puzzles, genannt Tangram. Auf der rechten Seite befindet sich eine Anordnung der Teile. Finden Sie eine Möglichkeit, die Teile so anzuordnen.*

- Three versions of the task are given in Figures B.6, B.7 and B.8.

- **Note:** In the experimental settings, Tangram pieces can be translated and rotated in the plane. However, 3D rotations are not allowed. This restriction was intentionally introduced to make the third version of the task (Figure B.8) unsolvable, i.e., the parallelogram cannot be positioned appropriately to represent "the flame of the candle" unless it is rotated in 3D.

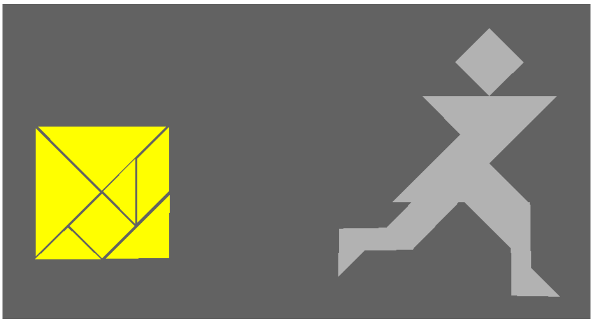


Figure B.6: The first version of the task "Tangram puzzle".

B.5 Grid Puzzle

- **Task description:** The Grid puzzle consists of an $[n \text{ x } n]$ grid that contains $n^2 - 1$ tiles numbered from 1 to $n^2 - 1$, and the n^2th place is empty. Before play begins the tiles are scrambled. The objective of this puzzle is to unscramble the tiles to get them into consecutive order. Subject is allowed only to make moves which slide tiles into the empty space.

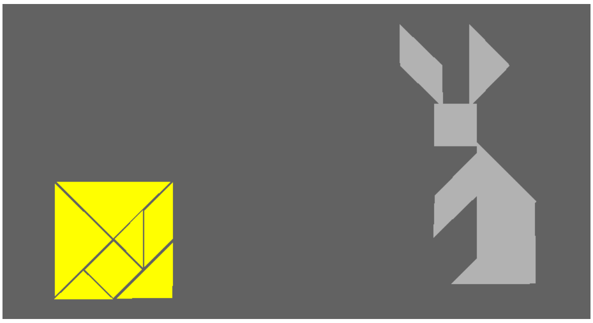


Figure B.7: The second version of the task ”Tangram puzzle”.

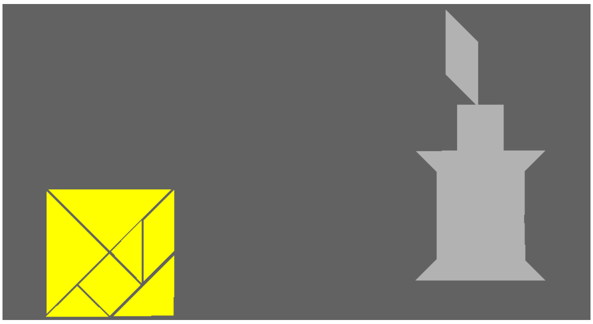
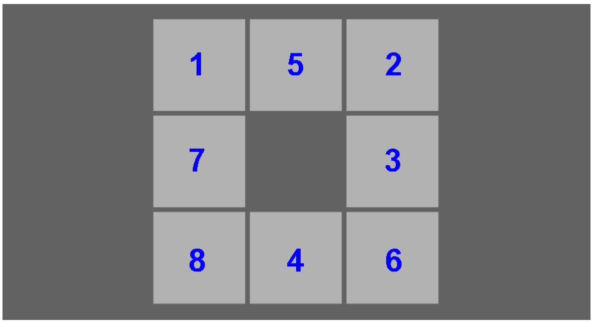


Figure B.8: The third version of the task ”Tangram puzzle”.

- **Introduction made by the wizard:** *Bitte ordnen Sie die Quadrate in richtiger, aufsteigender Reichfolge an. Sie dürfen nur einzelne Teile in das leere Feld bewegen.*

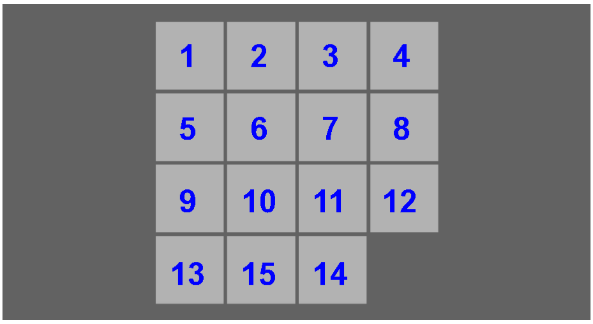
- The task was given in three versions. The first and the third versions of the task are given in Figures B.9 and B.10. In the second version of the task, the captured subject's image is mapped into 3x3 grid. The right bottom part of the image is discarded. Other parts are scrambled in the same way as in the first version of the task (Figure B.9). To protect the identity of the subjects, the picture of this version of the task is not provided.

- **Note:** Third version of the task (Figure B.10) is unsolvable.



1	5	2
7		3
8	4	6

Figure B.9: The first version of the task ”Grid puzzle”.



1	2	3	4
5	6	7	8
9	10	11	12
13	15	14	

Figure B.10: The third version of the task ”Grid puzzle”.

B.6 Tower of Hanoi Puzzle

- **Task description:** The puzzle consists of three pegs and several disks of different sizes. At the start of the game, the disks are stacked in order of size on the leftmost peg. The goal of the puzzle is to move the entire stack to the rightmost peg according to the following rules: only one disk can be moved at a time, each move consists of taking the upper disk from one of the pegs and placing it onto another peg, and no disk may be placed on top of a smaller disk.

- **Introduction made by the wizard:** *Ihre Aufgabe ist es, die Scheiben vom linken Turm auf den rechten Turm zu bewegen. Sie dürfen immer nur jeweils die oberste Scheibe eins Turms auf einmal bewegen, und eine größere Scheibe darf nicht auf eine kleinere Scheibe gelegt werden.*

- The two versions of the task are given in Figures B.11 and B.12.

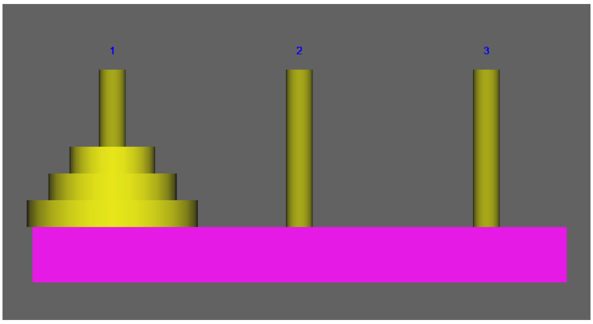


Figure B.11: The first version of the task "Tower of Hanoi puzzle".

B.7 The Three Jugs Problem

- **Task description:** There is an eight liter jug of water, and two empty jugs—a three liter jug, and a five liter jug. The objective of this task is to measure exactly four liter of water. Subject is allowed only to pour water from one jug to another jug.

- **Introduction made by the wizard:** *Sie sehen einen mit Wasser gefüllten Acht-Liter-Wassereimer, und weiterhin zwei leere Drei- und*

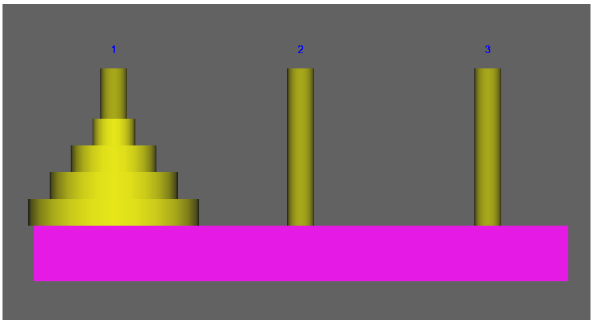


Figure B.12: The second version of the task "Tower of Hanoi puzzle".

Fünf-Liter-Wassereimer. Ihre Aufgabe ist es, genau vier Liter abzumessen. Sie können entweder einen Eimer komplett umschütten, oder einen anderen bis zum Rand füllen, und den Rest im Eimer lassen.

- This task is given in Figure B.13.

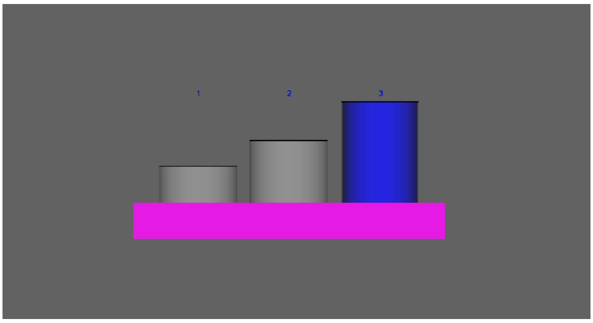
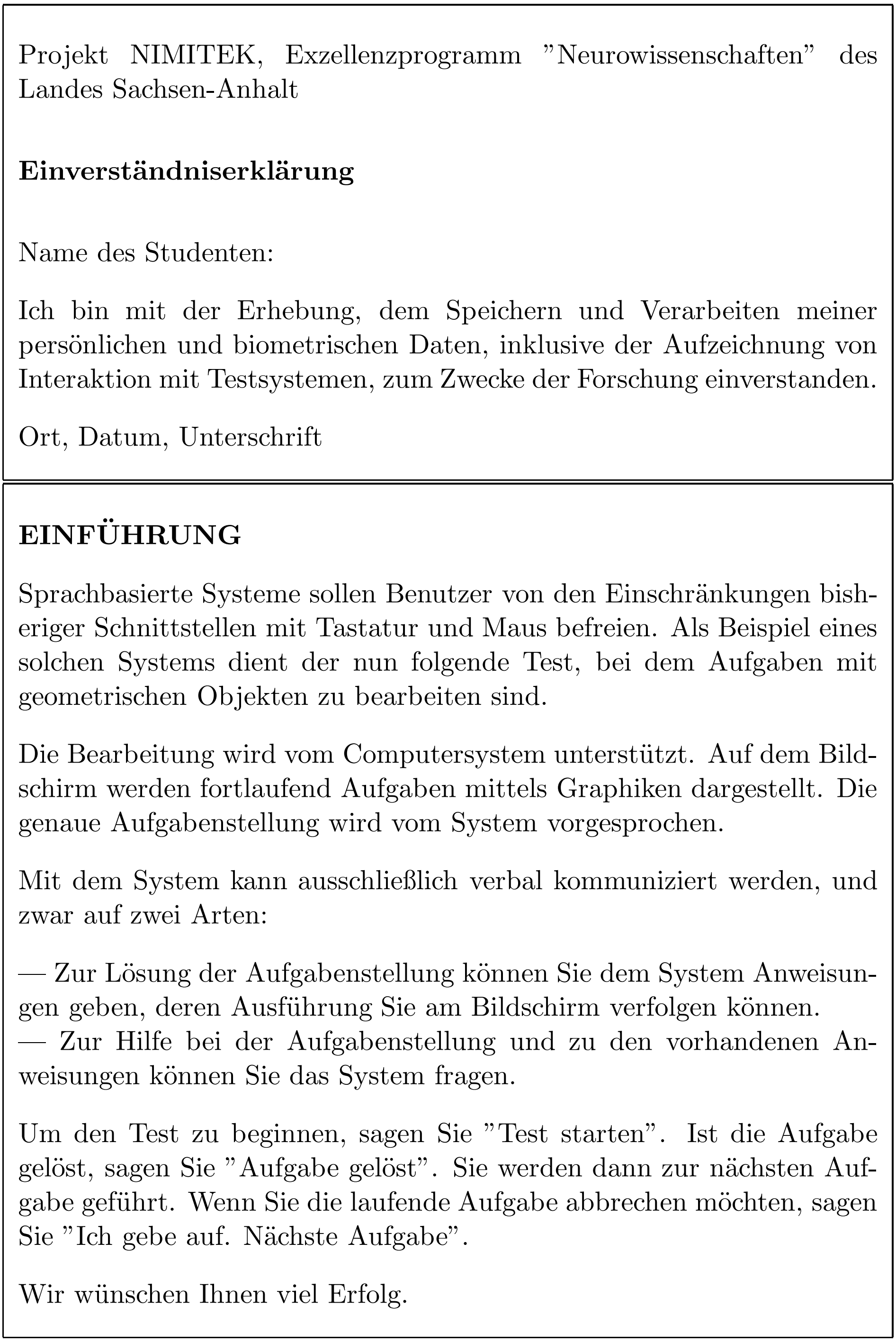


Figure B.13: The task "The Three Jugs Problem".

B.8 Consent to Participate in Research

Before the start of the test, the subjects gave the consent to participate in research. They were also given a short written description of the test. The text of these documents are given in Figure B.14.



Projekt NIMITEK, Exzellenzprogramm ”Neurowissenschaften” des Landes Sachsen-Anhalt

Einverständniserklärung

Name des Studenten:

Ich bin mit der Erhebung, dem Speichern und Verarbeiten meiner persönlichen und biometrischen Daten, inklusive der Aufzeichnung von Interaktion mit Testsystemen, zum Zwecke der Forschung einverstanden.

Ort, Datum, Unterschrift

EINFÜHRUNG

Sprachbasierte Systeme sollen Benutzer von den Einschränkungen bisheriger Schnittstellen mit Tastatur und Maus befreien. Als Beispiel eines solchen Systems dient der nun folgende Test, bei dem Aufgaben mit geometrischen Objekten zu bearbeiten sind.

Die Bearbeitung wird vom Computersystem unterstützt. Auf dem Bildschirm werden fortlaufend Aufgaben mittels Graphiken dargestellt. Die genaue Aufgabenstellung wird vom System vorgesprochen.

Mit dem System kann ausschließlich verbal kommuniziert werden, und zwar auf zwei Arten:

— Zur Lösung der Aufgabenstellung können Sie dem System Anweisungen geben, deren Ausführung Sie am Bildschirm verfolgen können.
— Zur Hilfe bei der Aufgabenstellung und zu den vorhandenen Anweisungen können Sie das System fragen.

Um den Test zu beginnen, sagen Sie ”Test starten”. Ist die Aufgabe gelöst, sagen Sie ”Aufgabe gelöst”. Sie werden dann zur nächsten Aufgabe geführt. Wenn Sie die laufende Aufgabe abbrechen möchten, sagen Sie ”Ich gebe auf. Nächste Aufgabe”.

Wir wünschen Ihnen viel Erfolg.

Figure B.14: Consent to participate in research and description of the test.

Bibliography

V. Aharonson and N. Amir. Emotion Elicitation in a Computerized Gambling Game. In *Proceedings of the 3rd International Conference on Speech Prosody 2006*, pages 179–183, Dresden, Germany, 2006.

J. Alexandersson and T. Becker. Overlay as the Basic Operation for Discourse Processing in a Multimodal Dialogue System. In *The 2nd IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, Seattle, 2001.

J. Alexandersson and N. Reithinger. Learning Dialogue Structures from a Corpus. In *Proceedings of EuroSpeech-97*, pages 2231–2235, Rhodes, 1997.

J. Alexandersson, B. Buschbeck-Wolf, T. Fujinami, M. Kipp, S. Koch, E. Maier, N. Reithinger, B. Schmitz, and M. Siegel. Dialogue Acts in VERBMOBIL-2, Second Edition, Verbmobil-Report 226. Technical report, DFKI Saarbrücken, Universität Stuttgart, Technische Universität Berlin, Universität des Saarlandes, 1998. Public Report.

J. Alexandersson, R. Engel, M. Kipp, S. Koch, U. Küssner, N. Reithinger, and M. Stede. Modeling Negotiation Dialogs. In W. Wahlster, editor, *Verbmobil: Foundations of Speech-to-Speech Translation*, pages 441–451. Springer-Verlag Berlin Heidelberg, 2000.

J. Allen. Collaborative Dialogue Agents. Invited speech held in the framework of the Fourth International Workshop on Human-Computer Conversation (Bellagio 2008), 2008. URL `http://www.companions-project.org/downloads/Companions_Bellagio08_Allen_Slides.pdf`.

A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth. Desperately Seeking Emotions: Actors, Wizards, and Human beings. In *Proceedings of the ISCA Workshop on Speech and Emotion: A Conceptual Framework for Research*, pages 195–200, 2000.

A. Batliner, C. Hacker, S. Steidl, E. Nöth, S. D'Arcy, M. Russell, and M. Wong. "You stupid tin box"—children interacting with the AIBO robot: A cross-linguistic emotional speech corpus. In *Proceedings of the 4th International Conference of Language Resources and Evaluation LREC 2004*, pages 171–174, 2004.

A. Batliner, S. Steidl, C. Hacker, E. Nöth, and H. Niemann. Private Emotions vs. Social Interaction—towards New Dimensions in Research on Emotion. In *Adapting the Interaction Style to Affective Factors (Workshop on Adapting the Interaction Style to Affective Factors, 10th International Conference on User Modelling)*, Edinburgh, 2005. (8 pages, no pagination).

A. Batliner, S. Biersack, and S. Steidl. The Prosody of Pet Robot Directed Speech: Evidence from Children. In *Proceedings of the 3rd International Conference on Speech Prosody 2006*, pages 1–4, Dresden, Germany, 2006.

A. Batliner, S. Steidl, C. Hacker, and E. Nöth. Private emotions versus social interaction: a data-driven approach towards analysing emotion in speech. *User Modelling and User-Adapted Interaction—The Journal of Personalization Research*, 18(1–2):175–206, 2008.

E. Bilange. An approach to oral dialogue modelling. In M. Taylor, F. Néel, and D. Bouwhuis, editors, *The Structure of Multimodal Dialoque II*, pages 189–205. John Benjamins Publishing Company Philadelphia/Amsterdam, 2000.

D. Bohus and A. Rudnicky. Sorry, I Didn't Catch That! An Investigation of Non-Understanding Errors and Recovery Strategies. In L. Dybkjær and W. Minker, editors, *Recent Trends in Discourse and Dialogue*, volume 39 of *Text, Speech and Language Technology*, pages 123–154. Springer, 2008. ISBN 978-1-4020-6820-1.

M. Brewer. Research Design and Issues of Validity. In H. Reis and C. Judd, editors, *Handbook of Research Methods in Social and Personality Psychology*, pages 3–16. Cambridge University Press, 2000.

W. Burleson. *Affective Learning Companions: Strategies for Empathetic Agents with Real-Time Multimodal Affective Sensing to Foster Meta-Cognitive and Meta-Affective Approaches to Learning, Motivation, and Perseverance.* PhD thesis, Massachusetts Institute of Technology (MIT), 2006.

N. Campbell. On the Structure of Spoken Language. In *Proceedings of the 3rd International Conference on Speech Prosody 2006*, Dresden, Germany, 2006.

J. Carbonell. Requirements for robust natural language interfaces: the LanguageCraft^{TM} and XCALIBUR experiences. In *Proceedings of the COLING-86*, pages 162–163, Washington, D.C., USA, 1986.

R. Catizone, A. Setzer, and Y. Wilks. Deliverable D5.1 State of the Art in Dialogue Management. Technical report, The COMIC Project (IST-2001-32311), 2002. URL `http://www.hcrc.ed.ac.uk/comic/documents/deliverables/D5-1Final.pdf`. Public report.

A. R. Damasio. *Descartes' Error: Emotion, Reason and the Human Brain.* Gosset/Putnam Press, New York, 1994.

L. Devillers and J.-C. Martin. Coding emotional events in audiovisual corpora. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, 2008.

L. Devillers, R. Cowie, J.-C. Martin, E. Douglas-Cowie, S. Abrilian, and M. McRorie. Real life emotions in french and english tv video clips: an integrated annotation protocol combining continuous and discrete approaches. In *Proceedings of the Fifth Inernational Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, 2006.

E. Douglas-Cowie and WP5 members. Preliminary plans for exemplars: Databases. Technical report, The HUMAINE Association, 2004. URL `http://emotion-research.net/projects/humaine/deliverables/D5c.pdf`. Public report.

E. Douglas-Cowie, R. Cowie, and M. Schröder. A new emotion database: Considerations, sources and scope. In *Proceedings of the ISCA Workshop on Speech and Emotion*, pages 39–44, Northern Ireland, 2000.

E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. McRorie, J.-C. Martin, L. Devillers, S. Abrilian, A. Batliner, N. Amir, and K. Karpouzis. The humaine database: Addressing the collection and annotation of naturalistic and induced emotional data. In *Proceedings of the Second International Conference on Affective Computing and Intelligent Interaction (ACII'2007)*, pages 488–500, Lisbon, Portugal, 2007.

E. Douglas-Cowie, R. Cowie, C. Cox, N. Amir, and D. Heylen. The sensitive artificial listener: an induction technique for generating emotionally coloured conversation. In *Proceedings of the Second International Workshop on Corpora for Research on Emotion and Affect (satellite of LREC'08)*, pages 1–4, Marrakech, Morocco, 2008.

M. Fék, N. Audibert, J. Szabó, A. Rilliard, G. Németh, and V. Aubergé. Multimodal spontaneous expressive speech corpus for hungarian. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, 2008.

K. Forbes-Riley, D. Litman, S. Silliman, and A. Purandare. Uncertainty corpus: Resource to study user affect in complex spoken dialogue systems. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, pages (5 pages, no pagination), Marrakech, Morocco, 2008a.

K. Forbes-Riley, D. Litman, and M. Rotaru. Responding to student uncertainty during computer tutoring: An experimental evaluation. In *Proceedings 9th International Conference on Intelligent Tutoring Systems (ITS)*, pages 60–69, Montreal, Canada, 2008b.

N. Fraser and G. Gilbert. Simulating speech systems. *Computer Speech and Language*, 5:81–99, 1991.

M. Gnjatović. An Array-Based Data Model for Tabment Selection. Master's thesis, Faculty of Computer Science, Otto-von-Guericke-University Magdeburg, 2005.

M. Gnjatović and D. Rösner. Gathering Corpora of Affected Speech in Human-Machine Interaction: Refinement of the Wizard-of-Oz Technique. In *Proceedings of the International Symposium on Linguistic Patterns in Spontaneous Speech (LPSS 2006)*, pages 55–66, Academia Sinica, Taipei, Taiwan, 2006.

M. Gnjatović and D. Rösner. An approach to processing of user's commands in human-machine interaction. In *Proceedings of the 3rd Language and Technology Conference (LTC'07)*, pages 152–156, Adam Mickiewicz University, Poznan, Poland, 2007a.

M. Gnjatović and D. Rösner. A Dialogue Strategy for Supporting the User in Spoken Human-Machine Interaction. In *Proceedings of the XII International Conference "Speech and Computer" (SPECOM'2007)*, pages 708–713, Moscow State Linguistic University, Moscow, Russia, 2007b.

M. Gnjatović and D. Rösner. Processing Dialogue-Based Data in the UIMA Framework. In *UIMA Workshop at the GLDV 2007*, Tübingen, Germany, 2007c. URL http://incubator.apache.org/uima/downloads/gldv/gldv07-uima-gnjatovic.pdf. Web published.

M. Gnjatović and D. Rösner. Emotion Adaptive Dialogue Management in Human-Machine Interaction. In *Proceedings of the 19th European Meetings on Cybernetics and Systems Research (EMCSR 2008)*, pages 567–572, Austrian Society for Cybernetic Studies, Vienna, Austria, 2008a.

M. Gnjatović and D. Rösner. On the Role of the NIMITEK Corpus in Developing an Emotion Adaptive Spoken Dialogue System. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, 2008b. URL http://www.lrec-conf.org/proceedings/lrec2008/pdf/149_paper.pdf. Web published.

M. Gnjatović and D. Rösner. The NIMITEK Corpus of Affected Behavior in Human-Machine Interaction. In *Proceedings of the Second International Workshop on EMOTION (satellite of LREC'08): Corpora for Research on Emotion and Affect*, pages 5–8, Marrakech, Morocco, 2008c. URL http://www.lrec-conf.org/proceedings/lrec2008/workshops/W2_Proceedings.pdf. Web published.

M. Gnjatović and D. Rösner. Adaptive Dialogue Management in the NIMITEK Prototype System. In *Proceedings of the 4th IEEE Tutorial and Research Workshop Perception and Interactive Technologies for Speech-Based Systems (PIT'08)*, volume 5078 of *Lecture Notes in Computer Science*, pages 14–25, Kloster Irsee, Germany, 2008d. Springer. ISBN 978-3-540-69368-0.

M. Gnjatović, M. Kunze, X. Zhang, J. Frommer, and D. Rösner. Linguistic Expression of Emotion in Human-Machine Interaction: The NIMITEK Corpus as a Research Tool. In *Proceedings of the Fourth International Workshop on Human-Computer Conversation (Bellagio 2008)*, Bellagio, Italy, 2008e. URL http://www.companions-project.org/downloads/Companions_Bellagio08_Gnjatovic_Paper.pdf. Web published.

B. Grosz and C. Sidner. Attention, Intentions, and the Structure of Discourse. *Computational Linguistics*, 12(3):175–204, 1986.

R. Guindon. How to Interface to Advisory Systems? Users Request Help With a Very Simple Language. In *Proceedings of ACM Conf. on Computer Human Interaction (CHI88)*, pages 191–196, Washington, D.C., USA, 1988.

M. Halliday. *An Introduction to Functional Grammar.* Edward Arnold, London New York, Second edition, 1994.

D. Heckmann, T. Schwartz, B. Brandherm, S. M., and M. von Wilamowitz-Moellendorff. Gumo—the general user model ontology. In *Proceedings of the 10th International Conference on User Modeling (UM'2005)*, pages 428–432, Edinburgh, UK, 2005. Springer-Verlag Berlin Heidelberg, LNAI 3538.

D. Heckmann, E. Schwarzkopf, J. Mori, D. Dengler, and A. Kröner. The user model and context ontology gumo revisited for future web 2.0 extensions. In *Proceedings of the International Workshop on Contexts and Ontologies: Representation and Reasoning (C&O:RR)*, pages 37–46, Roskilde, Denmark, 2007.

C.-H. Lee. Fundamentals and Technical Challenges in Automatic Speech Recognition. In *Proceedings of the XII International Conference "Speech and Computer" (SPECOM'2007)*, pages 25–44, Moscow State Linguistic University, Moscow, Russia, 2007.

D. Litman and S. Silliman. Itspoke: An intelligent tutoring spoken dialogue system. In *HLT-NAACL 2004: Demonstration Papers*, pages 5–8, Boston, Massachusetts, USA, 2004. Association for Computational Linguistics.

E. Lucas. La tour de Hanoï et la question du Tonkin. In *Récréations Mathématiques*, pages 285–286. Reprinted by Blanchard, Paris, France, 1959. Original published by Gauthier-Villars, Paris, 1884.

D. Luzzati. A Dynamic Dialogue Model for Human-Machine Communication. In M. Taylor, F. Néel, and D. Bouwhuis, editors, *The Structure of Multimodal Dialoque II*, pages 207–221. John Benjamins Publishing Company Philadelphia/Amsterdam, 2000.

J. R. Martin. Types of Structure: Deconstructing Notions of Constituency in Clause and Text. In E. H. Hovy and D. R. Scott, editors, *Computational and Conversational Discourse, Burning Issues An Interdisciplinary Account*, pages 39–66. Springer-Verlag Berlin Heidelberg, 1996.

I. Mazzotta, F. de Rosis, and V. Carofiglio. Portia: a user-adapted persuasion system in the healthy-eating domain. *IEEE Intelligent Systems*, 22 (6):42–51, 2007.

M. McTear. Handling Miscommunication: Why Bother? In L. Dybkjær and W. Minker, editors, *Recent Trends in Discourse and Dialogue*, volume 39 of *Text, Speech and Language Technology*, pages 101–122. Springer, 2008. ISBN 978-1-4020-6820-1.

M. McTear, I. O'Neill, P. Hanna, and X. Liu. Handling errors and determining confirmation strategies - an object-based approach. *Speech Communication*, 45(3):249–269, 2005.

R. Niese, A. Al-Hamadi, and B. Michaelis. A Novel Method for 3D Face Detection and Normalization. *Journal of Multimedia*, 2(5):1–12, 2007.

I. O'Neill, P. Hanna, X. Liu, and M. McTear. The Queen's communicator: An object-oriented dialogue manager. In *Proceedings of the of EuroSpeech 2003*, Geneva, 2003.

H. Pirker and G. Loderer. I said "two ti-ckets": How toTalk to a Deaf Wizard. In *Proceedings of the ESCA Workshop on Dialogue and Prosody*, pages 181–185, 1999.

I. Rahwan and P. McBurney. Argumentation Technology. *EEE Intelligent Systems*, 22(6):21–23, November/December 2007.

E. Roulet. On the Structure of Conversation as Negotiation. In J. L. Mey, H. Parret, and J. Verschueren, editors, *(On) Searle on conversation*, pages 91–99. John Benjamins Publishing Company, Philadelphia/Amsterdam, 1992.

E. A. Schegloff. Sequencing in conversational openings. *American Anthropologist*, 70:1075–1095, 1968.

J. Searle. Conversation. In J. L. Mey, H. Parret, and J. Verschueren, editors, *(On) Searle on conversation*, pages 7–29. John Benjamins Publishing Company, Philadelphia/Amsterdam, 1992a.

J. Searle. Conversation Reconsidered. In J. L. Mey, H. Parret, and J. Verschueren, editors, *(On) Searle on conversation*, pages 137–147. John Benjamins Publishing Company, Philadelphia/Amsterdam, 1992b.

G. Skantze. Exploring human error recovery strategies: implications for spoken dialogue systems. *Speech Communication*, 45(3):325–341, 2008a.

G. Skantze. Galatea: A Discourse Modeller Supporting Concept-Level Error Handling in Spoken Dialogue Systems. In L. Dybkjær and W. Minker, editors, *Recent Trends in Discourse and Dialogue*, volume 39 of *Text, Speech and Language Technology*, pages 155–189. Springer, 2008b. ISBN 978-1-4020-6820-1.

B. Vlasenko, B. Schuller, A. Wendemuth, and G. Rigoll. Frame vs. Turn-Level: Emotion Recognition from Speech Considering Static and Dynamic Processing. In *Proceedings of 2nd International Conference on Affective Computing and Intelligent Interaction (ACII 2007)*, pages 139–147, Lisbon, Portugal, 2007.

W. Watt. Habitability. *American Documentation*, 19:338–351, 1968.

A. Wendemuth, J. Braun, B. Michaelis, F. Ohl, D. Rösner, H. Scheich, and R. Warnemünde. Neurobiologically inspired, multimodal intention recognition for technical communication systems (nimitek). In *Proceedings of the 4th IEEE Tutorial and Research Workshop Perception and Interactive Technologies for Speech-Based Systems (PIT'08)*, volume 5078 of *Lecture Notes in Computer Science*, pages 141–144, Kloster Irsee, Germany, 2008. Springer. ISBN 978-3-540-69368-0.

B. Wendt and H. Scheich. The "Magdeburger Prosodie-Korpus". In *Proceedings of the Speech Prosody 2002 Conference*, pages 699–701, Laboratoire Parole et Langage, Aix-en-Provence, 2002.

Y. Wilks, R. Catizone, and M. Turunen. Dialogue Management. Technical report, COMPANIONS Consortium: State of the Art Papers, 2006. URL http://www.companions-project.org/downloads/Companions_SoA2_Dialogue_Management.pdf. Public report.

C.-H. Wu, Z.-J. Chuang, and Y.-C. Lin. Emotion recognition from text using semantic labels and separable mixture models. *ACM Transactions on Asian Language Information Processing (TALIP)*, 5(2):165–183, 2006.

W. Xu, B. Xu, T. Huang, and H. Xia. Bridging the Gap Between Dialogue Management and Dialogue Models. In *Proceedings of the Third SIGdial Workshop on Discourse and Dialogue*, pages 201–210, Philadelphia, 2002.