In man-machine communication, as in everyday dialogs, intelligent dialog behavior of a natural language (NL) system involves the ability of the system to shift its cognitive perspective in order to take the role of the dialog partner and to simulate his or her dialog behavior. One particular way in which a system can anticipate a user’s responses is to make use of the system’s own comprehension (and perhaps generation) capabilities, temporarily taking the role of the user and simulating his or her behavior. In the user modeling literature, such cases are characterized as so called *anticipation feedback loops (AFL)*.

To date, almost all implemented systems that have employed anticipation feedback have used a limited part of the system to realize a *local* AFL. The are few examples of systems that use a *global* AFL. With a global AFL, a large part of the system’s own understanding capabilities is used to anticipate the user’s responses. **Chapter 1** motivates the thesis and introduces the issues raised by global anticipation feedback. **Chapter 2** presents an interdisciplinary view of the concept of anticipation. The similarity to the concept of *role-taking* in developmental psychology is reflected in the theoretical framework. Relations to psycholinguistics are also pointed out.

In this thesis, we model AFLs with decision trees: A system $S$ has to choose among several possible dialog moves $m_1 \ldots m_n$ that will have some effect on the user $U$ (cf. the left-hand side of Figure 8.1). Each $m_i$ has some immediate degree of appeal for $S$, which can be conceived of as a utility $U_m(m_i)$. But instead of selecting the move with the highest $U_m(m_i)$, $S$ anticipates the response $r_i$ that $U$ is likely to make to each $m_i$; and each $r_i$ is itself associated with a utility $U_r(r_i)$. $S$ chooses the move with the highest total utility $U_m(m_i) + U_r(r_i)$. An AFL can be invoked in the step where $S$ anticipates $U$’s response $r_i$. The point of doing so is that the determination of $U_r(r_i)$ in addition to $U_m(m_i)$ may affect $S$’s choice of a move. This simplicity and generality is, however, associated with a number of limitations and challenges, as the following issues raised by global AFLs show:

1. **Within-dialog transmutability.** A system that uses a global AFL must be able to take the role of the other participant in the type of dialog it conducts. By contrast, a local AFL presupposes only that the system be able to do some part of the processing required for the other role; and this common processing may involve a generic subtask, such as syntactic analysis, which is relatively independent of any particular dialog role. For human beings, transmutability is often given, because people learn to take many different roles in dialogs in the course of their everyday experience. (For example, even a professional salesperson often has the opportunity to act as a customer.) But systems that employ user modeling techniques are typically designed to play a particular role. It may therefore require a considerable additional investment to enable them to switch to the role of their dialog partner.
2. **Communication between system instantiations.** The dialog system has to be able to invoke itself in another system instantiation, in the role of the dialog partner, without interfering with its workings in its original role.

3. **Uncertainty about factors that determine the responses of the dialog partner.** The decision tree in the left-hand side of Figure 8.1 presupposes that $S$ can predict $U$’s response to any given move $m_i$ with certainty. The more general case is shown in the right-hand side of the figure: When considering a move $m_i$, $S$ can at best narrow the possible responses of $U$ down to a set $\{r_{ij}\}$. In the case of a global AFL, this uncertainty is due to the fact that $U$’s response will be influenced by some factors that are not entirely known to $S$. In other words, $S$ does not know exactly how to pretend to be the user. The question arises of how $S$ can best deal with this uncertainty.

4. **Resource-adaptive Anticipation.** In general, it is relatively time-consuming for a dialog system to anticipate the complete processing by the user of a possible dialog move by the system. The question therefore arises of how the computational cost of using global AFLs can be minimized.

The present thesis explores the relatively uncharted area of global anticipation feedback with the help of the listed issues. These questions are investigated within the NL dialog system PRACMA (Jameson et al., 1994; 1995), which we use here as a testbed.

**Flexible Architecture and Transmutability**

Chapter 3 introduces some basic architectures for NL systems and then focuses on the CHANNELS architecture, which uses techniques from distributed AI and from concurrent object-oriented programming. (PRACMA is implemented within this architecture.) In CHANNELS (Cooperating Heterogeneous Agents for a Natural-Language System), the system’s modules, realized as agents, interact cooperatively through a communication-act-based protocol that governs the exchange of messages. Each message is characterized by attributes including: the sender, the recipient(s), the type of communication act, the mode...
of communication (synchronous or asynchronous), and the actual content of the message. The communication acts (INFORM, ASK, and REPLY) define the nature of the interaction among the agents. For instance, an ASK requests the recipient(s) to send information back to the originator of the message, while an INFORM passes information from one module to another. Messages communicate information between a sender agent and a receiver agent either synchronously or asynchronously. The communication is synchronous if the sender requires a response before continuing processing; until a response is received, it remains in the state waiting. With asynchronous communication, the sender can engage in further processing before receiving a response. The agents run concurrently as simulated processes over a local network using PVM (Geist et al., 1994) and ICE (Amtrup, 1994). For the distribution of the computation of the agents over a local network and the realization of the communication between various instantiations of the PRACMA system, the CHANNELS architecture supports external agents. (An external agent is a virtual agent within a PRACMA instantiation to which the agents within the system can refer and send messages but which is actually located in another instantiation of the system on the same or on another machine.) These external agents are used for the realization of global anticipation feedback.

Most natural language systems are restricted to cooperative dialogs, whereas PRACMA (PRocessing Arguments between Controversially-Minded Actors) models partly noncooperative dialogs, which are widely represented in everyday situations. PRACMA models dialogs in which a person (to be called the seller) is trying to sell his or her used car to a potential buyer. PRACMA is able to take the role of either the seller or the buyer within its dialog situation. This ability to switch roles can be seen as a particular variant of the property of dialog systems that Wahlster and Kobsa (1989, p. 30) define as transmutability: the property of being adaptable to applications that differ with respect to “dialog type, user type, and intended system behavior”. This within-dialog transmutability in PRACMA enables the system to realize various types of AFLs—in particular, truly global AFLs, in which the system consults a complete instantiation of itself in order to anticipate the dialog partner’s responses.

The goals of the two dialog participants in PRACMA’s example domain conflict to a certain degree: The buyer wants to get the best possible information on which to base a decision about the car, whereas the seller would like to sell the car, whether or not it is really suitable for the buyer. When S is the seller, this conflict increases the importance of anticipation feedback for S, because it increases the range of utility that U’s responses can have for S. For example, if U decides to ask about an attribute of the car, it makes a big difference to a noncooperative seller whether this attribute happens to be one on which the car rates highly or poorly; for a cooperative seller, this difference would not be so important. (The car is modeled in terms of six evaluation dimensions: Reliability, Economy, Safety, Environment, Comfort and Sportiness.)

PRACMA uses a simple local AFL when taking the role of a seller who is not able or willing to devote much attention to the dialog. This AFL is local in that the only part of the system’s comprehension capabilities that S makes use of is S’s capability to derive the evaluative implications of a given comment about a car. The use of a global AFL in PRACMA becomes necessary when S tries to anticipate a more complex response by U: What U’s next
dialog move will be if $S$ makes a given dialog move. It can be important for $S$ to anticipate $U$’s next move. For example, if $U$ asks a question, it might concern some topic that $S$ would prefer to avoid (e.g., an attribute with respect to which the car is weak); if this seems likely, $S$ should consider not making the comment that it originally intended to make.

Realization of Global Anticipation Feedback

The job of choosing the system’s next dialog move is divided hierarchically between the PRACMA agents DIALOG PLANNER and COMMENT AND QUESTION HANDLER (nicknamed CQH). The DIALOG PLANNER, an incremental planner, decides what type of move to make. In doing so, it takes into account a variety of factors, including the dialog history (stored in the agent PRAGMATIC DIALOG MEMORY) and various motivational parameters (stored in the agent EGO). Once it has decided on a particular type of move, it asks CQH to choose a specific move of that type.

A global AFL cannot be realized as a query to one of agents that make up the system, as it requires an invocation of the entire system. Therefore, a special agent, the GLOBAL ANTICIPATOR (nicknamed GAF for Global Anticipation Feedback) maintains a subordinate instantiation of PRACMA in a separate COMMONLISP image which may be located on another computer. GAF initializes this instantiation to take the other dialog role. This instantiation does not engage in any direct interaction with the user but rather responds to inputs from the GAF that controls it. (Recall that the CHANNELS architecture provides external agents that make possible communication between different instantiations.) This subordinate image contains instantiations of all of PRACMA’s agents, including GAF. The GAF in the main instantiation accepts and responds to queries from agents that require global anticipation feedback. One task of GAF is to ensure that the subordinate instantiation constitutes as realistic a model of the actual user as is possible given the information available to the main instantiation. To accomplish this, GAF regularly queries the agents EVALUATION HANDLER and DOMAIN BELIEF HANDLER in the main instantiation, asking them for their assessments of $U$’s evaluation criteria and knowledge; the resulting estimates are used to initialize the corresponding agents in the subordinate instantiation.

Anticipation of the Buyer’s Processing

In Chapter 4 we present a straightforward use of global anticipation feedback for the prediction of the reaction of the buyer. In this approach, GAF uses an instantiation of $B$ (the buyer) with exactly the same evaluation standards, domain knowledge, and dialog strategies that $S$ herself would have if she were buying a used car. In this approach, (called situational role taking by Higgins (1981)), a communicator predicts the responses of the communication partner by asking, “What would I do, think, feel, see, etc., if I were in that situation?” The use of this approach may be based on assumed similarity, where the communicator does not even consider the possibility that he or she may differ in important ways from the dialog partner; or on inferred similarity, a considered judgment that such differences are not important enough to have to be taken into account.
$S$ anticipates $B$'s next dialog move by temporarily pretending that it is the buyer in the dialog. More concretely, $S$ consults a simultaneously active instantiation of PRACMA in which the system is taking the role of the buyer. $S$ uses the strategy described in the left-hand side of Figure 8.1. For each potential utterance $m_i$ whose utility $U_m(m_i)$ is above a given threshold $\delta$, $m_i$ is sent to the subordinate instantiation (where the system takes the role of the buyer) to anticipate the potential reaction $r_i$ of $B$. The resulting utility $U_r(r_i)$ is then added to the already computed utility $U_m(m_i)$. $S$ eventually chooses the move with the highest expected overall utility $U_m(m_i) + U_r(r_i)$.

**Taking Uncertainty into Account**

In Chapter 5 we present an approach in which $S$ takes into account his or her uncertainty about the factors that determine the responses of the dialog partner.

The algorithms used in the simple variant have presupposed that, if $S$ takes the trouble to use a global AFL, $S$ will always anticipate $U$'s response correctly. This assumption is less realistic than the conceptualization shown in the right-hand side of Figure 8.1. But the question remains: How can global anticipation, given a possible move $m_i$, return not just a single anticipated response $r_{i1}$, but rather a set of possible responses $\{r_{i1} \ldots r_{in}\}$? This question is difficult to answer in a general way. But within the framework presented here, the problem is manageable if $S$ considers only other responses of the same type as the most likely response $r_{i1}$. The basic idea is to exploit the way in which CQH chooses moves of a given type, namely by evaluating all reasonable moves of that type. Although this algorithm has been discussed so far only with respect to its use in the main instantiation, it is of course also used in the subordinate instantiation, when $U$ is being simulated. For example, when the subordinate instantiation, in the role of the buyer, chooses a specific question to ask, the CQH of the subordinate instantiation first considers all questions that have some relevance to the current dialog focus and then chooses the one with the highest UTILITY-OF-MOVE. A consequence is that when the subordinate instantiation has produced a move $r_{i1}$ for $U$ as a response to the move $m_i$ by $S$, GAF can ask CQH which moves it considered that had a UTILITY-OF-MOVE that was almost as high as that for $r_{i1}$. The assumption underlying this query is the following one: The moves that rated almost as high as $r_{i1}$ for the simulated $U$ represent the most likely alternative hypotheses about how $U$ will respond to $m_i$.

By applying the described strategy, $S$ becomes more cautious, asking not just “What response would I make in this situation?” but rather “What responses would I consider making in this situation?” The resulting $N$ responses that would rate highest for $S$ if she took the role of $B$ can be viewed as the $N$ most likely responses for $B$ to make. The assumption underlying this runner-up strategy is: “Even though I have a good deal of uncertainty about the $B$ I’m talking with, he is probably not so different from me that he would choose a move that was not among my $N$ most preferred moves”. (Actually these $N$ moves are those whose overall utility exceeds a given threshold $\epsilon$.) The higher $N$ (i.e. the lower $\epsilon$) is set, the more likely it is that $B$’s actual response will be anticipated. If $S$ checks that she could in turn respond satisfactorily to all of these $N$ possible responses, she is relatively unlikely to overlook potential problems. (Note that anticipating a set of possible responses
is in itself no more time-consuming than anticipating a single one, since CQH always has
to consider several moves. The list of almost-chosen moves is already available to the CQH
in the subordinate instantiation as a by-product of its processing.)

**Anticipation of the Seller's Processing**

A buyer in general knows that the seller is maintaining a model of his interests and
knowledge, and the buyer may try to anticipate how this model—and the seller’s dialog
moves based on this model—will be affected by particular behaviors of B. Chapter 6 shows
how PRACMA, when taking the role of the buyer, anticipates the behavior of the seller and
the seller’s modeling of the buyer.

An important topic is this approach is the notion of an *impression*. In the role of the
seller, S maintains a model of the buyer. This model contains the impressions that the buyer
has about the car. The seller chooses his dialog contributions so that these impressions of
the buyer are altered. These impressions are estimates of B’s interests in the six evaluation
dimensions. To anticipate the reactions of the seller, S simulates the processing of the seller
and estimates the seller’s impressions about the buyer’s interests in the evaluation dimen-
sions before and after an utterance. (Recall that six evaluation dimensions are considered by
PRACMA: Reliability, Economy, Safety, Environment, Comfort and Sportiness.) For each
potential utterance, S in the role of the buyer anticipates whether this utterance increases the
estimate of his interest in the main evaluation dimension that the seller ascribes to the buyer.
The utility is a function of the resulting impression shift concerning the buyer’s interests. S
ultimately chooses his dialog contribution according to its utility.

**Resource-Adaptive Anticipation**

Because global anticipation feedback is computationally expensive, a system must be
selective in applying it. Chapter 7 deals with resource-adaptive anticipation and discusses
several possible types of taking into account resource limitations. The presented approaches
can be classified according to criteria whether, how, when global anticipation should be used
and how much look-ahead is required.

The strategies of global anticipation presuppose that all branches of the considered de-
cision trees are to be processed completely. But if a system is willing to sacrifice some
decision quality according to the available system resources, computation can be done more
selectively. For example, the satisficing approach shows how the search for a move can be
terminated as soon as one as an acceptable overall utility has been found—i.e. the overall
utility of an intended move exceeds some threshold. A special case of the satisficing strategy
is the iterative approach: The system does not generate all possible moves $m_i$ at once but
rather generates and evaluates them one by one, using the result of the evaluations $U_r(r_{ij})$
to guide the generation of the next move $m_{i+1}$.

The approach of post-utterance anticipation decides under time pressure when to anticipate
according to the expected probability that the move selected without global anticipation is
adequate or not. Depending on this probability, the system might decide to use global anticipation before generating an utterance or to refrain from using global anticipation beforehand and use it after the generation of the selected move while the dialog partner is analyzing the system utterance. If necessary, the system initiates additional utterances to compensate for the undesired consequences of the verbalized utterance.

In the approach of anticipation according to the value of computation, the system first has to decide whether it is worthwhile to anticipate before the verbalization of a move or not. The system first selects the two best moves without use of global anticipation. It then considers the expected improvement in utility compared to choosing a move without applying global anticipation to these moves. According to the variance of the distribution of the utilities the system chooses the move which is expected to be the best or it decides to actually anticipate in the cases where there is a good deal of uncertainty.

The next approach deals which the restriction of the use of global anticipation feedback to a late stage in the utterance planning process. For example, in PRACMA, when deciding what type of dialog move to make next, the dialog planner often asks the module for selecting utterances whether it is possible to make a worthwhile move of a given type. But even if the answer is positive, the planner may end up choosing a different type of dialog move, since other criteria are also relevant. Because this type of query by the DIALOG PLANNER comes frequently, it would be impractical for the selection module to invoke the global anticipation module (perhaps repeatedly) every time it answers such a query. Instead, it simply checks whether there is some move of the type in question that is acceptable. It is only when (and if) the planner subsequently asks the selection module actually to select a move of this type that it takes the trouble to invoke the anticipation module. When it does so, it may of course discover that all of the possible moves rate poorly. In such cases the system's behavior is similar to that of a person who begins to say something and then has second thoughts about the wisdom of doing so. The occasional appearance of this phenomenon seems to be a necessary consequence of the limited time that the system can spend anticipating the user's responses during the early planning of a dialog contribution.

The next strategy handles the selective updating of the subordinate instantiation on the basis of estimates of the user's knowledge, evaluation criteria, and other characteristics. Performing this updating frequently can be not only time-consuming but also wasteful. For example, only a small part of the updates may actually have any effect on the anticipation of $U$'s next move. A simplified approach is to do the updating only occasionally—or even only once, at the beginning of a dialog, on the basis of the initially available information about the user.

The next approach deals with the minimizing look-ahead. In principle $S$ could expand the nodes of the used decision trees so as to look at least several moves ahead as is done, e.g., in game-playing programs. (Note that the system views the user not as an adversary but simply as a cause of events that have a limited degree of predictability.) Another type of look-ahead can be achieved if global anticipation is allowed to occur within the subordinate instantiation: When anticipating $U$'s next response, $S$ considers how $U$ will anticipate $S$'s subsequent move, etc. An important limitation of both of these types of look-ahead is their relatively high computational cost. For example, to extend a decision tree beginning with
one of the right-most nodes, $S$ has to go through the whole process of generating possible moves, a process which can involve all of the agents which make up the system. Note also that as the tree gets deeper, the additional expansions become less worthwhile, as they concern dialog moves which are increasingly unlikely ever to occur. One reasonable approach is to make the amount of look-ahead dependent on (a) the resources available to the system and (b) the assessed importance of correct anticipation.

The last approach deals with the anticipation assuming that the dialog partner is under time pressure. The simulation of this time pressure involves e.g. the restriction of the selection process in the subordinate instantiation to moves whose topics are sufficiently close to the selected item in the dialog focus.

**Conclusion**

Chapter 8 looks back on the themes developed in the previous chapters to see what has been contributed: a general framework for realizing and comparing variants of global anticipation feedback; one solution to the problem of realizing communication between system instantiations; a manageable approach to the handling of uncertainty about the user; a bidirectional anticipation ability; and some approaches to resource-adaptive anticipation. The chapter points out what remains to be done as future work.