1. Introduction

The study of language by means of computers has become fairly mainstream since the turn of the millenium. On the one hand, very large amounts of language data have become available which can be analyzed in various ways, ranging from simple word searches to parsing and statistical modeling as well as signal processing and video analysis. On the other hand, the increasing demand for natural language applications such as machine translation and speech interfaces is pushing development towards steady increases in quality and usefulness. Technical advances in building interactive systems have brought forward the need for studying language not only as a means of conveying simple commands and propositions, but also for expressing affections, controlling the interaction, and creating a shared understanding and social bonds among the interlocutors.

Computational pragmatics refers to the study by means of computers of how language is used in context. This is notably done by building models of how meaning is created in context. The need for elaborate computational pragmatic models grows bigger as more advanced interactive systems are being designed such as spoken dialog systems, embodied conversational agents, and robots: these applications require rich interaction capabilities for verbal and non-verbal communication. Surprisingly, perhaps, few systems for machine translation have substantial pragmatic knowledge: the current paradigm consists of statistical training on aligned samples in the source and target languages. In contrast, computer systems aiming at
conducted a *spoken or written dialog* face challenges which can only be addressed with a well-developed pragmatic component. For that reason, much research in pragmatics has been in the context of developing *intelligent interactive agents*, and the present chapter will focus on this research.

This chapter is structured along the different types of pragmatic components and various degrees of contextual information needed to model intelligent agents: starting from the sentential context and moving on to two-party interactions and finally to situated interaction with embodied agents. We start with discourse relations and speech acts viewed in terms of intentions, inferencing and the stepwise planning that computers require. We review early work in computational modeling related to *chat-bots*, pattern-matching interactive agents which react to a given input without much semantic or pragmatic knowledge about the partner's intentions or the interactive situation.

We move on to *spoken dialog systems* which prepare for a multi-turn interaction with the user, and introduce different dialog models such as Grosz and Sidner's (1986) model, intention-based models, and conversation analysis. We also discuss the role of cooperation and shared context, and issues related to the planning and generation of discourse. We continue with *embodied conversational agents* which may include models for various multimodal communication channels (speech, gesture, facial displays, body posture) for the understanding and presenting of information. Next, we turn to *situated conversational agents*, which take the interaction modeling into the three-dimensional world and engage the user into interaction in a particular environment.

## 2. Sentences in context

As a matter of course, many of the issues that are raised in discourse and conversation analysis also turn up in the computer processing of natural language. For a correct interpretation of an utterance, the hearer needs to take into account the context of the communication, including the situational context, the participants' knowledge of the domain and their intentions, and the links between the current utterance and those that precede it.
The importance of background knowledge, and especially the ability of the language users to bring that knowledge to bear on the process of language comprehension and production, has been widely recognized in a number of approaches that have been traditionally labeled *pragmatic*, such as speech act theory (Austin 1962; Searle 1975, 1979), discourse processing (Grosz & Sidner 1986) and dialog management (Jokinen & McTear 2009). The added value of a computational approach is the requirement that knowledge of whatever kind be dealt with in a systematic and objective way in order for a program to work. Ideally a well-constructed program functions as a model that provides an accurate generalisation of the particular phenomenon and allows for experimentation by simulation. *Artificial Intelligence* (AI) research provides a whole range of techniques to support the design of a knowledge base, its formal representation, and the implementation of the mechanisms to apply that knowledge (Barr & Feigenbaum 1981; Luger & Stubblefield 1983; Winston 1992; Russell & Norvig 2009).

2.1 *Intentions, inferencing and planning*

Grice's theory of meaning (Grice 1975) describes communication in terms of mental states, i.e. beliefs and intentions of the participants, in relation to the speaker's intention to convey meaning to the listener. The ultimate goal of a speaker's communicative act is to influence the listener's mental state so that the listener's future actions and attitudes will accord with the intentions of the speaker. Successful communication thus requires that the listener recognizes the speaker's purpose in using language, adopts the goal (at least temporarily for the time of the interaction), and plans her own actions so as to assist the speaker to achieve the underlying goal, or at least so as not to prevent the speaker from achieving her goal. This presupposes that the partner is cooperative.

The view of language as action is rooted in the *speech act* theory formulated by Austin (1962) and Searle (1975, 1979). Language does not only serve to express propositions, but also to perform communicative actions. Since people do not act randomly but plan their actions to achieve certain goals, each communicative act is tied to some underlying plan that the speaker intends to follow, and intends to communicate to the partner. Cohen and Perrault (1979) and Allen and
Perrault (1980) view linguistic behavior as goal driven planning: the speaker plans an utterance in order to achieve a communicative goal, while the listener's aim is to infer that goal from the linguistic form.

This view articulates the specific goals and plans a speaker may have in using language and it has been instrumental in AI approaches to dialog planning. A specific example, adapted from Wilensky (1981), involves the goal and associated plan for asking, including the preconditions that need to be satisfied for the plan to be successful:

| Goal: | X wants to find out P from Y |
| Act:  | Ask question to Y |
| Preconditions: | 1. X is near to Y  
| | 2. Y knows P  
| | 3. Y wants to tell P to X |
| Result: | Y tells P to X |

The goal-directed and plan-based view on linguistic behavior has been widely adopted and is the prevailing view of communication in the current AI community (e.g. Russell & Norvig 2009). The main research topics have evolved from the earlier explicit logic-philosophical work on dialogs and intentions (e.g. Kobsa 1989; Carberry 1989; Cohen et al. 1990) to statistical and machine-learning approaches to dialog policy and user simulation (Scheffler & Young 2000; Schatzmann et al. 2006; Williams & Young 2007).

A single dialog contribution can be multifunctional, and a single speech act can span several contributions (Allwood 1978). For instance, *We'll take the twin room then* can simultaneously function as a statement, an acceptance, and a request, while also dealing with task-level information (e.g. reserving a twin room) and meta-level dialog control issues (e.g. closing the topic). On the other hand, requests like wanting to reserve a room may be completed over a sequence of contributions where the individual utterances serve to establish a common ground. Appropriate references are also often negotiated so that the participants agree that they identify the same referent (Clark & Wilkes-Gibbs 1986).

The underlying conceptual representation in a cognitive psychology approach is the source of further reasoning. In order to arrive at an answer to a question, it is often insufficient to convert the question directly into a database query, but
frequently inferences have to be drawn. This is acknowledged by Lehnert (1980) in his discussion of QUALM, the reasoning component of a question-answering system using Schank’s Conceptual Dependency and scripts approaches (Schank 1972; Schank & Abelson 1977). Answering a question involves various manipulations of the conceptual structure underlying the question, such as inferring what entities are actually referred to by a wh-word.

Consider for example the question *Who wasn’t at the party?* Probably a considerable part of the world’s population was not present, but a reasonable inference is that the question refers only to invitees who did not show up. Besides a correct identification and conceptualization of mentioned events, knowledge is also needed about inferred events that probably occurred but were not mentioned (such as an invitation), failed expectations, etc. These can be made on the basis of knowledge about stereotypical situations as well as general knowledge about how people achieve goals and what sorts of goals they try to achieve (Wilensky 1981).

2.2 Discourse relations

Much on discourse processing has been restricted to fairly simple discourse types with a strong tie to a particular task. For instance, Grosz (1977) studies interactions in which an expert instructs an apprentice on how to assemble an air compressor. Grosz shows that it is possible to formulate focusing heuristics because the task restricts what is talked about: the structure of the task is mirrored in the dialog structure. Similarly, Carberry (1989) establishes focusing heuristics that rely on the expectations of possible shifts of focus constrained by the underlying task-related plan in an information-seeking dialog.

The cooperative nature of the discourse types studied also aids in establishing coherence relations. For instance, McKeown (1985) expanded the focus rules designed by Sidner (1983, 1985) in her computational work on discourse generation. Her approach is based on the observation that people follow certain standard patterns or schemas of discourse generation (see below) for attaining discourse goals. These patterns are helpful for the listener or reader in establishing the thread of discourse and are thus an aid in understanding (Carberry 1989). *Relevance theory* (Sperber & Wilson 1986) is based on the view that the speakers intend their utterances to bear meaningful inferential links with each other, i.e. to bear relevance with respect to the previous discourse. Listeners can then interpret
the discourse by identifying the most informative relations without too deep inferential steps, thus aiming for maximal informational benefit while minimizing the cost of inferencing.

Following Gricean cooperation principles, Reiter (1990) discusses how the choice of a referring expression influences the implicatures that the user is able to draw. For instance, when selecting *shark* instead of *dangerous fish*, the speaker relies on the partner's knowledge that sharks are dangerous animals, whereas preferring *dangerous fish* to *shark*, the speaker implies that the type of the fish is unknown but that the fish is known to be dangerous.

While much work has been done in generating coherent texts, the generation of coherent dialogs has been studied in less detail. In text linguistics, (e.g. Halliday & Hasan 1976; Beaugrande & Dressler 1981), the distinction between *cohesion* and *coherence* is usually drawn. The former means surface level ties between textual elements (like pronominalization), the latter inferable links between ideas and objects in the text. In dialog management, links between objects have been modeled with the help of topic and focus (e.g. McKeown 1985; Grosz & Sidner 1986; McCoy & Cheng 1991), while the ideas in the successive discourse segments can be captured by rhetorical relations like *Elaboration, Parallel* and *Contrast* (Mann & Thompson 1987, 1988), conjunctive relations (Halliday & Hasan 1976), or coherence relations (Hobbs 1979). Recognition of the speaker intentions can also be understood as the basic coherence mechanism: discourse is coherent if the underlying purpose is shared by the participants and if each contribution contributes to achieving this purpose (e.g. Cohen & Perrault 1979; Cohen & Levesque 1990). However, as pointed out by Grosz and Sidner (1986), the speaker's intentions must be distinguished from the thematic coherence or the focusing structure, since the same intentional structure can give rise to different focusing structures in different discourses.

Hobbs (1979) argues that discourse coherence is a deeper notion than "discourse just being about some set of entities", i.e. it involves reasoning. For instance, the discourse *John took a train from Paris to Istanbul. He likes spinach* would be normally considered incoherent, even though it is about John. Hobbse explains that coherence stems from the agent's need to be understood, which drives the agent to seek for a suitable coherence relation. In the TACITUS project, Hobbs et al. (1990) put forward an approach that uses *abductive* reasoning to interpret
discourse. The interpretation is related to the logical form of the text and a minimal explanation is needed of why the text would be true in its context. For instance, the literal meaning of the sentence *Boston office called* does not make sense, since offices cannot make phone calls, but the sentence can be understood by finding an explanation through abduction: the person working in the office had made a phone call. In a similar manner, a discourse can be found coherent if an inferential link between the sentences explains their logical relation, for instance, in the example above, if Istanbul is known for its spinach dishes and John’s liking spinach is the reason for his traveling to Istanbul.

A number of computational models have been proposed for the comprehension of discourse, but they tend to ignore its complexities (Gamham 1995). Several approaches tend to treat discourse as a product, for which stereotypical knowledge structures can be implemented. Reichman (1985), for instance, proposes a kind of grammar for discourse, consisting of formal discourse rules and an ATN formalism for analyzing discourse. This contrasts with the approach taken in conversation analysis research, in which dialog is seen as an interactional achievement.

### 3. Interactive situations

As already discussed, the participants engaged in communication try to convey a particular meaning so as to achieve a particular goal. The addressee gives meaning to the utterances by interpreting them in the given context according to the general communicative principles and one’s underlying intentions. Communication is also cooperative activity, and the participants provide feedback to each other in order to resolve ambiguities and misunderstandings, and to affirm that they have understood what the partner said. Moreover, dialog is mixed-initiative, in that all partners can start a new topic, for instance, and thus influence the progress of the dialog. By starting a new topic, the speaker can introduce issues that support the speaker’s own goals and intentions, but at the same time they have to be aware of the partner’s feedback and ready to respond in an appropriate way to the partner’s topics. Early AI research has experimented with models that embody some of this communicative capacity.
3.1 Natural language interfaces

Possibly the first system capable of carrying out an extended dialog in a given context was SHRDLU (Winograd 1972). The system was grounded in an artificial world of colored blocks, pyramids and a box, all of which can be manipulated by a simulated robot arm which could be operated through commands, as shown in the following extract of a dialog:

Person: PICK UP A BIG RED BLOCK.
Computer: OK. (does it)
Person: GRASP THE PYRAMID.
Computer: I DON'T UNDERSTAND WHICH PYRAMID YOU MEAN.
Person: FIND A BLOCK WHICH IS TALLER THAN THE ONE YOU ARE HOLDING AND PUT IT INTO THE BOX.
Computer: BY "IT", I ASSUME YOU MEAN THE BLOCK WHICH IS TALLER THAN THE ONE I AM HOLDING.
Computer: OK. (does it)
Person: WHAT DOES THE BOX CONTAIN?
Computer: THE BLUE PYRAMID AND THE BLUE BLOCK.
Person: WHAT IS THE PYRAMID SUPPORTED BY?
Computer: THE BOX.
Person: HOW MANY BLOCKS ARE NOT IN THE BOX?
Computer: FOUR OF THEM.
Person: IS AT LEAST ONE OF THEM NARROWER THAN THE ONE WHICH I TOLD YOU TO PICK UP?
Computer: YES, THE RED CUBE.

As is apparent from this example, SHRDLU was capable of interpreting nontrivial linguistic expressions in a discourse context as well as planning and reasoning in a changing quasi-physical world. One of the key innovations that SHRDLU embryonically established is the systematic representation and use of two interlocking kinds of knowledge: linguistic knowledge, pertaining to the syntax and the semantics of the language, and world knowledge, including information on what exists in the world and how it works. It was made clear that understanding and producing natural language requires substantial amounts of knowledge.
about the *domain* of discourse, as well as reasoning about the *structure* of the discourse. It also became clear that handling the knowledge needed in a toy world is one thing, but designing a system for the real world is another. Scaling up a system is a formidable challenge which involves integrating vast amounts of knowledge.

### 3.2 Question answering

In the 1960s, a number of computer programs rooted in early AI techniques enabled people to communicate with computers in written natural language. For instance, **Baseball** (Green et al. 1963) allows a user to query the system about the dates, locations, teams and scores of baseball games. **Baseball** answers questions like *Who beat the Yankees on July 4?* The system is an example of a natural language interface to a database: the user enters natural language text which is transformed by the system into a formal query; on the output side, the retrieved data are presented to the user in the form of a natural language sentence. Other examples include **ProtosynthesexI** (Simmons et al. 1964), a question-answering system that has as its domain the contents of a children’s encyclopedia. and **Student** (Bobrow 1968), a system that is able to find sets of equations in algebraic story problems and solve them. **Lunar** (Woods 1972) answers question about lunar rocks; it is one of the first systems using a sophisticated linguistic formalism, **Augmented Transition Networks** (ATNs), to process natural language (Woods 1970).

A modern, powerful variant of a question answering system is the **Watson DeepQA** system by IBM, that has successfully participated in *Jeopardy* games, in which the tasks of questioning and answering are reversed (Ferruci et al. 2010). For instance, to the clue *Aeolic, spoken in ancient times, was a dialect of this*, the system correctly responded with the spoken question *What is Ancient Greek?* Unlike the earlier mentioned systems, Watson is not limited to a narrow domain, but extracts all kinds of knowledge from a vast number of sources. Still, like the above-mentioned systems, it essentially deals with isolated question-answer pairs and does not handle more complex communicative situations. Real life information systems must be capable of maintaining a sustained dialog with clarifications, confirmations, etc. (e.g. Aust et al. 1995), and maintain mixed-initiative dialogs (Chu-Carroll 2000).
3.3 Conversational agents

Conversational systems or conversational agents aim at maintaining interaction that is more human-like chatting. Even rather simplistic techniques can be used to give a fairly convincing appearance of intelligence in casual conversation. A well-known example is ELIZA (Weizenbaum 1966), a program that casually converses with its user in English by accepting any input and responding more or less appropriately. A particularly popular version of the program is the simulation of a Rogerian psychotherapist conducting an initial interview with a patient. The dialog between ELIZA and its patients looks sophisticated at first sight, as can be seen in the following extract:

**ELIZA:** EARLIER YOU SAID YOU WERE HAVING TROUBLE WITH THE COMPUTER
**User:** that’s right and it hasn’t got any better
**ELIZA:** WHAT DOES THAT SUGGEST TO YOU?
**User:** that computers are no use
**ELIZA:** HOW WOULD YOU REACT IF MACHINES TOOK OVER?
**User:** I think it would be a disaster
**ELIZA:** DOES ANYONE ELSE THINK IT WOULD BE A DISASTER?
**User:** Yes quite a lot of people
**ELIZA:** QUITE A LOT OF PEOPLE
**User:** that’s right

This dialog was taken from a later version of ELIZA (McTear 1987). The program does not perform any serious linguistic analysis of the input: it scans the input for particular keywords and patterns and then generates an automatic response based on pattern matching. For instance, the phrase I think triggers a response asking if also other people have the same opinion. Also, it uses particular phrases that are meant to show interest by simply reflecting the patient’s words or to prompt the patient for further input (as for instance, What does that suggest to you?). The particular communicative setting, where it is generally the patient who talks and the therapist prompts the patient, enables the program to carry on this superficial conversation strategy. Simple repetition may be enough to suggest that the dialog partner is listening empathically.

ELIZA and related systems were based on sets of patterns and heuristics (rules of thumb) to guide a conversation in a restricted subset of the language. However, the fact that even psychotherapists entertained the idea that a refined ELIZA might
be able to substitute a human psychotherapist (Weizenbaum 1984) suggests that artificial language users might ultimately pass the Turing test (Turing 1950, see Section 6.1). Modern versions of Eliza, so called conversation bots or chatbots, such as Alice, have indeed become more sophisticated and, for instance, Alice won the Loebner Prize as “the most human computer” at the annual Turing Test contests in 2000 and 2001 (Wallace 2009). Nevertheless, more well-founded pragmatic, semantic and syntactic models are necessary to address a wider range of communicative needs in goal-directed dialogs, as will be discussed below.

From the interaction modeling point of view, however, Eliza and Alice possess no real understanding of utterances or conversational principles, and the user soon realizes that the dialog leads nowhere. What is missing is the mechanism to take care of the speakers’ intentions and the coherence of the dialog (Jokinen 2009). In the 1980s and 1990s several projects started to integrate knowledge of the language processing and dialog management into a unified system that would produce interaction in a natural and robust manner. Several collaborative projects focused on dialog building systems; for instance, the European project Plus (Pragmatics-based Language Understanding System) (Black et al. 1991) aimed at the challenging goal of robust natural language dialog system, by exploiting both the knowledge of pragmatic linguistic phenomena, such as interpretation with respect to context, and the power of inference tools derived from non-linguistic problems, in order to provide reasoning-based robustness in natural language understanding.

In the remainder of this chapter, we will give an overview of how computational approaches to pragmatics contribute to modeling the analysis and generation of utterances (including multimodal expressions) in context. Discourse representation, text planning, and dialog management are crucial parts of this endeavor. Also a domain model and world knowledge which ground communication in a representation of reality, and a user model that infers the user’s knowledge, goals and plans, are important components of such computational models.

4. Dialog modeling

The early computational models for dialog were formalised as dialog grammars, following the rule-based grammars of language. The grammar-based approach is grounded in the observation that dialogs have a number of sequencing
regularities (questions are followed by answers, proposals by acceptances, etc.),
and in the assumption that these regularities can be captured by a dialog gram­
mar (Sinclair & Coulthard 1975; Reichman 1985; Moeschler 1989). A dialog
grammar defines dialog units such as moves, exchanges and segments, as well
as their possible combinations, and the rules can range from straightforward
re-writing rules to those with elaborated embedded constructions. Popularity of
dialog grammars is mainly due to their formal properties which allow efficient
implementation via finite state automata or context free grammars. Nowadays
the grammar approach appears in the form of script-based dialog management
where the script describes the overall dialog structure and the state transitions
implement dialog rules.

Another approach to the pragmatic language studies is intention-based mod­
eling, which has its roots in the development of speech act theory (Austin 1962;
Searle 1979) and especially in the formalization of speech acts as planning opera­
tors. Speech acts allowed interaction to be modeled via the speaker's intentions,
and they thus provided a way to take into account the dynamic aspect of language
usage. In the development of dialog systems, an important milestone was the obser­
vation that speech acts could be operationalized with the help of planning opera­
tors, commonly used in the AI modeling of actions (Cohen & Perrault 1979; Allen
& Perrault 1980). Planning operators define a set of preconditions that need to be
fulfilled in order for the operator to apply, and a set of postconditions that describe
what the world will be like after the application of the operator (see an example in
Section 2.1).

The original concept of speech act was also extended with dialog information
so that they describe the whole dialog state and include contextual information
besides the action itself. Dialog acts (also termed communicative acts, conversa­
tional acts) have been an important means to model interaction, and much research
has been conducted on the segmentation, recognition, and prediction of dialog acts
(see e.g. Stolcke et al. 2000). For instance, dynamic interpretation theory (DIT)
defines dialog acts as a combination of the communicative function and semantic
content (Bunt 1990). The communicative function is an operation on the context,
and thus, when a sentence is uttered, its semantic content is expressed and also a
set of background assumptions is changed. Bunt (2006) proposes a multidimen­sional
description of dialog acts, since dialog is an activity in which participants
exchange information, while also monitoring the communicative process, social
obligations, etc. This work builds on the multi-layered approach in DAMSL (Core & Allen 1997) and has culminated in the ISO standardisation of dialog annotation (Bunt et al. 2012).

Dialog acts can also be seen as a special class of more general communicative acts. Cohen and Levesque (1990) consider specific utterance events in the context of the speaker’s and hearer’s mental states, and derive the different effects of the acts from general principles of rational agenthood and cooperative interaction.

One of the most famous discourse theories is developed by Grosz and Sidner (1986) who combine three independent, yet interconnected descriptive levels: linguistic level, intentional level, and attentional state. Each level uses its own organization that captures different aspects of discourse (and dialog) management. The linguistic level concerns linguistic expressions used to convey meaning; discourse cohesion is modelled on this level in terms of pronouns and other linguistic means to mark coreference. The speakers’ intentions and the discourse purposes are modeled in the intentional level, while the attentional state encodes focus spaces associated with discourse segments and the most salient objects in them. Together the levels determine the well-formedness of the discourse (and the dialog).

Conversation analysis (Goodwin 1981; Sacks et al. 1974) focuses on spontaneous everyday spoken interactions, and tries to describe the discourse on the basis of the participants’ conversational strategies to take turns, provide feedback, and build a shared context. Compared with the formal discourse analysis, also conversation analysis focuses on the structure of the dialog but rather than providing top-down rules that the speakers follow, it aims at explaining the structure being produced in the interaction itself. Conversation analysis points out that although dialogs have conventionalised opening and closing sequences, they are otherwise structured as they go on. A dialog structure can be recognized by external observation but is not an internal constraint of dialog management, and cannot be equated with the speakers’ conforming to particular predefined structural rules.

The more formalized AI approach on the one hand and the more empirical approach advocated by conversation analysts on the other hand have been conceived as irreconcilable opponents, though a fruitful interaction that may enrich the flexibility of human-computer interaction is advocated by Luff et al. (1990). Also, although typical conversation analysis research may be difficult to apply directly to the design of interactive systems, it has introduced a terminology about dialog
phenomena; examples of widely accepted terms are *adjacency pairs*, *turn taking*, *back-channeling*, *repairs*, and *opening and closing sequences*. It has also brought forward insights which are useful for spoken dialog design: utterances are not continuous stretches of speech but contain hesitations and pauses, they can overlap and provide non-verbal feedback, backchanneling. For instance, research on repairs has inspired computational models (McRoy & Hirst 1995; Heeman & Allen 1997; Krahmer et al. 1999), and the notion of shared construction of utterances and conversational meaning has influenced models of grounding and dialog cooperation (Clark & Wilkes-Gibbs 1986; Clark & Schaefer 1989).

An important part of the discourse and dialog modeling is to collect and analyze real data. An empirically and ecologically sound methodology for studying the multimodal nature of communication presupposes the existence of adequate dialog corpora, i.e. annotated collections of audiovisual recordings of the interaction of users in different communicative situations with other users or with systems. A number of such corpora have recently been developed in the projects such as AMI, CHILL, and HuMAINE (see Martin et al. 2007 for an overview) as well as in more application-oriented development of dialog systems. Much effort has also been put on defining annotations and coding manuals, and special initiatives and conferences focus on the collection, annotation, and analysis of dialog corpora in general.

### 4.1 Cooperation

Various kinds of conversational principles play a role in communication. These include, for instance, cooperation, politeness, rationality, and dialog strategies (Goffman 1970). Grice’s conversational maxims (Grice 1975) have been the standard conversational principles in dialog modeling and they provide best-practice guidelines for developing practical dialog systems. The maxims of *quality* (be truthful, don’t say anything for which you lack adequate evidence), *quantity* (be informative, but don’t make your contributions more informative than is required), *manner* (be brief and orderly, avoid obscurity and ambiguity), and *relevance*, have also been criticized as applying only to dialogs which deal with factual information. The maxims describe communication as an exchange of information where each utterance fulfils all the requirements exhaustively and at once. However, people often break the maxims but do not appear to be uncooperative, and dialogs also convey meaningful information about the participant’s mutual relation, contact and
dialog control, so that cooperation can actually be said to manifest itself in the ways in which these other aspects are interleaved with the information exchanges.

Intentional collaboration has been much studied in the frameworks of TEAMWORK (Cohen & Levesque 1991) and SHAREDPLANS (Grosz & Sidner 1990). Sometimes collaboration is distinguished from cooperation (e.g. Allwood 1976). Cooperation is a special attitude of the speakers, stemming from their rational agenthood, and it can be defined that rational, competent agents are engaged in ideal cooperation, if they:

1. are voluntarily striving to achieve the same purposes,
2. are ethically and cognitively considering each other in trying to achieve these purposes,
3. trust each other to act according to (1) and (2) unless they give each other explicit notice that they are not.

For instance, Jokinen (1996, 2009) talks about ideal cooperation as the starting point for successful human–computer communication, implemented as a series of constraints that restrict possible attitudes to be conveyed to the partner. Thus cooperation can be seen as a sign of naturalness of the interface, emerging from the system’s processing capabilities which maintain interaction with the user on the basis of relevant and truthful information. Although it contributes to the user’s positive experience of the interaction with the system, cooperation is different from the usability of the interface which refers to the general fitness of the system for the task.

The speakers are also bound by social commitments and obligations (Allwood 1976, 1994), and they need to ground their contributions (Clark & Schaeffer 1989). Traum and Allen (1994) consider various types of obligations which are then implemented as separate action rules, while Allwood et al. (2000) give examples of basic cooperative mechanisms of dialog.

Cooperation can manifest itself on several levels, from a tight collaboration in order to achieve a particular goal, to behavior patterns that occur simultaneously when the interlocutors interact. In psycholinguistic and social interaction studies such behavior where partners align their behavior with that of the partner is usually called alignment (Pickering & Garrod 2004), but also synchrony or mimicry or copying (e.g. Mancini et al. 2007). Copying of each other’s movements, gestures and body postures often occurs in conversations, and this kind of synchronous
behavior is a cue of cooperation between the participants building common ground. On the other hand, simultaneous *gaze turning* can also demonstrate the end of a topic or sequence (Jokinen & Parkson 2011; Jokinen et al. (in print)).

4.2 Shared context and grounding

In AI, *grounding* mainly concerns the agents’ understanding of the relation between language and the physical environment, and is regarded as one of the main research problems (Harnad 1989). In dialog management, grounding has been modeled via a presentation-acceptance cycle that the speakers are involved in (Clark & Wilkes-Gibbs 1986), and as a particular dialog act that the agents perform as part of the dialog plan (Traum 1994). Jokinen (1996, 2009) sees grounding as the speakers’ conformation to the general requirements of cooperative and coordinated interaction, following the approach in Allwood (1976).

**Constructive dialogue modeling** (CDM) (Jokinen 2009) describes communication as a fundamentally cooperative activity between rational agents: the participants react to contextual changes and push forward their goals in a joint activity. The model refers to the construction of mutual context via a repeated cycle of interpreting, evaluating and reacting to the partner’s contributions. It also refers to the ideal cooperative nature of communication which ‘constructs’ rather than ‘copies’ communicative activity into dialog contributions. The constructive epithet in CDM also refers to the construction of individual contributions via gradual specification of the communicative goals. The goals of the dialog participants can range from social affective ones such as ‘keeping the channel open’ to more specific, task-oriented ones such as planning a trip, providing information, or giving instructions. The participants interact by exchanging new information and by reasoning about the effect of the exchanged new information on the shared context. The agents thus construct a shared context in which the goals can be achieved, while success of the interaction depends on the cognitive and emotional impact of the actions on the hearer. In order to ensure maximal impact, the agents must provide new information as clearly as possible, by using suitable lexical items, prosody (pitch, stress, volume, speed), and non-verbal means (gestures, gazing, face expressions), while the partner must be aware of these means in order to integrate the intended meaning in the shared context. Important topics in interaction management are related to information presentation: planning and generation of appropriate responses, giving
feedback, and managing topic shifts. Studies of feedback in dialog thus contribute to our understanding of grounding and the different conversational strategies and styles of communication that are used in the construction of the shared knowledge.

4.3 Topic and new information in the presentation of information

Previous work in natural language generation has shown the importance of a topic in restricting the content of a contribution, determining appropriate referring expressions, and selecting surface level expressions. Topic and focus have also been used to account for thematically coherent discourse and thus to constrain what can be talked about next.

Grosz, Joshi and Weinstein (1995) distinguish between global and local coherence, as well as between global focus and centering, respectively. The former refers to the ways in which larger segments of discourse relate to each other, and accordingly, global focus refers to a set of entities that are relevant to the overall discourse. The latter deals with individual sentences and their combinations to larger discourse segments, and accordingly, centering refers to a more local focusing process which identifies a single entity as the most central one in an individual sentence. Each sentence can thus be associated with a single backward-looking center which encodes the notion of global focusing and a set of forward-looking centers which encodes the notion of centering. Strube and Hahn (1999) provide a revision of the centering model and propose that the ordering of discourse entities as possible forward-looking centers should be based on the functional information structure of the sentence rather than on the grammatical role of the entities.

McCoy and Cheng (1991) try to cover different types of focusing phenomena by referring to a model of the conceptual structure of the domain of discourse. They also introduce the notion of focus tree and argue that the tree structure is more flexible in managing focus shifts than a stack: instead of pushing and popping foci in a particular order onto/from the stack, the tree allows one to traverse the branches in different orders, and the coherence of the text can be determined on the basis of the distance of the focus nodes in the tree. The focus shifting rules are expressed in terms of the type of relationships which occur in the domain. In generation, they provide information about whether or not a topic shift is easy to process (and, similarly, whether or not the hearer will expect some kind of marker), and in analysis they help to decide on what sort of topic shifts are likely to occur.
Jokinen, Tanaka and Yokoo (1997) applied the idea of focus tree in spoken dialog processing. They make the distinctions between topical vs. non-topical informational units, i.e what the utterance is about vs. what is in background, and new vs. old information in the dialog context.

5. Generating discourse

Generating an extended piece of discourse involves some careful amount of planning. This complex task has conveniently been divided into two subtasks: deciding what to say and deciding how to say it. The former is sometimes called text planning or strategic generation (Thompson 1977), and involves choices regarding the selection and organization of information. The latter subtask is sometimes called linguistic realization or tactical generation, and involves lexical and syntactic choices for the computation of the linguistic form of the utterance. We will in this section be concerned only with the first subtask, text planning (or discourse planning). Overviews of research in natural language generation can be found in Kempen (1989) and McDonald (1992) and in volumes edited by Kempen (1987), Zock and Sabah (1988), Dale, Mellish and Zock (1990a), Paris, Swartout and Marn (1991), Dale, Hovy, Rösner and Stock (1992), and Horacek and Zock (1993). Computational models of discourse planning are reviewed from a psycholinguistic perspective by Andriessen, De Smedt and Zock (1995).

Generating discourse is a multiply constrained process in which various knowledge sources should be taken into account: knowledge of the domain of discourse, the situational context and past discourse, as well as knowledge about the interlocutor or reader. As will be indicated in Section 6.2, user models are an important part of discourse understanding, but the user’s plans and goals also play an important role in discourse generation. Detecting and using the user’s goal to provide an appropriate response has been the object of extensive research (e.g. Appelt 1985; Carberry 1983; McKeown 1985). Even though the use of a specific user model in the generation process has been questioned (Sparck Jones 1989, 1991), tailoring discourse to the user’s level of expertise and taking the user’s misunderstandings and other input into account are obviously fundamental communicative abilities (Kaplan 1983; McCoy 1988; Quilici 1989; Reiter 1990; Cawsey 1990; McKeown et al. 1990; Paris 1988; Chin 1989; Moore & Swartout 1991; Moore 1989; Jokinen & Kanto 2004).
In discourse generation, two approaches can be distinguished. The first approach can be characterized as conceptualizing generation as a kind of planning in the AI sense of the word, driven by the communicative goals of the speaker (Appelt 1985; Cohen & Perrault 1979). This means that at the strategic level, text is planned by reasoning about both the system’s and the user’s knowledge and beliefs, and that speech acts are meant to have a particular impact on the user’s beliefs and knowledge structures. It has been argued that this approach does not incorporate an explicit notion of textual coherence and hence will face serious problems when transcending the sentence level (Dale et al. 1990b; Moore & Swartout 1991).

The second approach emphasizes text structuring above the level of the sentence. To this end, McKeown (1985) and Paris and McKeown (1987) propose schemas, i.e. representations of stereotypical discourse strategies. For instance, McKeown proposes four schemas for describing objects: identification, constituency, attributive and contrastive schemas. Schemas mandate the content and the order of the clauses in paragraphs. However, they do not allow the dynamic reassembly of the basic parts into new paragraphs. In order to fix this drawback, McKeown et al. (1990) envisage generalizing schemas into a hierarchy of increasingly more general schemas. Recent trends, however, are aimed at learning surface patterns from corpora without explicit pragmatic reasoning (Ravichandran & Hovy 2002; Curto et al. 2011).

An alternative approach offering a more detailed and dynamic text structuring is rhetorical structure theory (RST; Mann & Thompson 1987, 1988). RST identifies basic rhetorical relations as the building blocks from which coherent paragraphs (and thus, ultimately also the schemas mentioned above) are composed. Using RST relations, text generation systems can dynamically put together paragraphs. Some examples of rhetorical relations between elements in texts are sequence, which is signaled by words like then, next, etc.; purpose, signaled by in order to; and alternative, signaled by or. Rhetorical relations are used at several levels of the text structure, down to the level of single clauses. Coherent discourse is attained if all parts of a text can be hierarchically structured by rhetorical relations. Thus, the relations in a stretch of discourse can be represented as a tree structure. The branches of the tree represent adjacent clauses and blocks of clauses between which a particular rhetorical relation holds.

Considerable research has been devoted to implementations of RST, i.e. the design of planning algorithms that dynamically assemble the elements of a text.
using RST relations (Hovy 1988, 1990, 1991; Cawsey 1990; Moore & Swartout 1991; Paris 1988, 1991; Scott & de Souza 1990). Other research deals with focus in discourse. Focusing refers to the way in which the writer guides the reader's attention throughout a text. This, in turn, has consequences for the correct interpretation of referring expressions, for example pronouns and definite noun phrases. McCoy and Cheng (1991) investigate how a discourse focus tree can be built parallel to the discourse structure tree to track the focus of attention throughout the text (see Section 4.3).

Interesting possibilities are created by *multimodal* human-computer interaction, i.e. using both the modalities of conversation and of graphic interaction such as pointing. McKeown et al. (1990) describes a system in which text and graphics are used for explanations. Claassen (1992) proposes Edward, a multimodal dialog system where graphic interaction in a model world is combined with natural language commands and questions. An added feature of Edward is the *continuous linguistic feedback generator* (CLFG) which gives natural language feedback on the user's actions.

6. **Situated interaction with intelligent agents**

In this section we review research where the modeling has been extended to take into account the context in which the interaction takes place. This means that the system needs to incorporate a good model of the user, as well as the whole communicative repertoire which includes not only verbal communication with semantically meaningful words and utterances but also the non-verbal communication including gesturing, body posture, and facial expressions.

6.1 **Intelligent agents**

The idea of human-computer interaction can be traced back to Alan Turing who in his influential article sketched a *Thinking Machine* (Turing 1950). This was an abstract machine which would operate according to complex algorithms and produce behavior which could be described as intelligent. As a sign of its intelligence, it could, for instance, be engaged in natural language conversations.
Turing proposed a test, later to be called the Turing test, to judge whether such a Thinking Machine would indeed be an intelligent entity, despite the fact that its observed behavior would be produced by following algorithmic, albeit complex, rules: if the machine's behavior could not be distinguished from that of humans, then it could be judged as intelligent.

We will not be concerned here with the question whether AI systems are really intelligent or what it means to really understand language; these are controversial philosophical issues (see e.g. Copeland 1993; Searle 1984; Pylyshyn 1984; Harnad 1989; Hayes et al. 1992). For our present purpose, we consider that AI has been defined as the study of intelligent interactive systems (Russell & Norvig 2009) where we see applications of computational pragmatics in such examples as intelligent cars and intelligent home applications. We concede, however, that ubiquitous computing (Weiser 1991) has brought the philosophical questions back, since we can ask if the autonomous applications in everyday life can be considered intelligent given that the human user cannot always have a direct control over the system reaction or its functioning, and how intelligence is related to communication between the human user and the application.

6.2 Modeling the user

User models are specific components designed for a better understanding of system users and their communicative needs. One of the first applications, the UNIX CONSULTANT (UC), a natural language program assisting beginning UNIX users, contains a component KNome which maintains a model of the user (Chin 1989; Wilensky et al. 1988). Some other systems which maintain user models are GRUNDY, which recommends novels to users in a library setting (Rich 1979), HAM-ANS, which assists a user in renting a room in a hotel (Jameson et al. 1980; Morik 1989), and XTRA, which acts as a tax advisor assisting the user in filling out his income tax form (Allgayer et al. 1989). In all these systems, a model of the user is part of a dialog component. However, it can readily be seen that user modeling is an important aspect of other applications such as intelligent computer aided instruction (Aleven et al. 2005), in which a teacher must monitor the knowledge of the student, and game playing, in which a system should take into account the perspective of its adversary in order to figure out what his plans and goals are.
In the context of dialog systems, a user model can be defined as a knowledge source which contains explicit assumptions on all aspects of the user that may be relevant to the dialog behavior of the system. Allen (1994) argues that regardless of whether the dialog is cooperative or not, a user model is the basis for intelligent dialog behavior. Among other things, it is required for identifying the objects talked about, identifying and analyzing non-literal meanings or indirect speech acts, determining the effects of the planned contribution on the listener, etc. As such, a user model is a crucial component of a dialog system: it provides important information for understanding the dialog partner as well as for producing an appropriate response.

The assumptions gathered in a user model must be separable by the system from the rest of the system's knowledge, and must be supplied to other components of the system which need them (Wahlster & Kobsa 1989). The intended separation between implicit and explicit models of the user stands in contrast to common practice in Human Computer Interaction, Software Ergonomics and Cognitive Engineering, where the designer of a software product has a typical user in mind, but the definition of that user is hidden in the system without being explicitly articulated; it can only be inferred from, for instance, the design of the user interface (Norman 1986; see also Norman & Draper 1986; Helander 1988). In a dialog system, the concepts user model and discourse model are closely related, but their exact relationship has been a matter of debate (Schuster et al. 1988).

There are various dimensions along which user models vary (Kass & Finin 1988). A first distinction is that between a canonical user model which accounts for all users and an individual user model which is specific for a single individual user. Canonical user models do not take into account the characteristics of individual users; once the latter come into play, a model must be explicitly maintained by the system and mechanisms must be provided for instantiating, updating and exploiting the model.

A second dimension concerns the long term versus short term user model. In the former, relatively stable or static characteristics of users are represented while in the latter specific interactional information (such as topics discussed, goals pursued, etc.) is stored. In this perspective the concept of a user model shows
close correspondence with that of a dialog model: a short term user model actually overlaps with a dialog model in that it records the specifics of the interaction (Rich 1989).

6.2.3 Construction of user profiles

Several paths can be followed in constructing a model of the user. First of all, the actual input of the user is a prime source from which his knowledge of the domain, as well as his plans and goals can be inferred. Secondly, this method can be combined with an approach that assumes a priori knowledge present in the system about types of users which is used as a basis for drawing the profile of an individual user.

The notion of a stereotype is useful for initiating a user model. Rich (1989) observes that facts about people tend to be interdependent in that particular traits of people appear to be clustered, forming stereotypes which each stand for a class of users. Rosch and Mervis (1975) use the term prototype as the denotation of such cognitive reference points. Stereotypes or prototypes enable a system to infer a whole set of user characteristics on the basis of a relatively small number of observations. For instance, in the user modeling component KNOME of the UNIX CONSULTANT (Chin 1989), users are characterized by four stereotypes: novice, beginner, intermediate, and expert, each of which represents an increasing mastery of the UNIX operating system. An individual user is an instantiation of one of the stereotypes and is assigned its default characteristics. In order to set up a model of an individual user, it is necessary to collect information from the user. This can be done in various ways:

- Users can be asked to classify themselves at the beginning of an interaction, as in GRUNDY (Rich 1989), or the REAL ESTATE AGENT (Mořík & Rollinger 1985).
- The user modeling component can be conceived in such a way that it ‘looks over the shoulder’ of the user and compares the user’s performance with that of its own built-in expert system. It is then possible to compare both and to deduce overlap and discrepancies between the user’s knowledge and the system’s in order to draw a map of the user’s knowledge (cf. Jokinen & Kanto 2004).
- User input can be analyzed to infer what knowledge about the domain it reveals, as in KNOME. A stereotype as introduced so far can now be seen as a set of assertions, irrespective of the way in which assertions are
represented in the system. Rich (1989) points out that stereotypes may be incorporated in a generalization hierarchy so that the mechanisms of (default) inheritance hold between the members in the hierarchy.

An important feature of the use of stereotypes is that a system can infer what users are likely to know or not, which user characteristics are likely to hold or not, based on only partial information about the user. In other words, the inference that a user belongs to a particular class defined by a stereotype enables the system to make a set of default inferences, which are plausible but defeasible. In order for these to work properly, uncertainty measures are associated with inferences, either as numerical values or as symbolic ratings of uncertainty (Rich 1983, 1989; Chin 1989). Moreover, in order to recover from contradicted inferences, not only the assertions about users are recorded, but also the justification of these assertions are noted, so that some form of truth maintenance (Doyle 1979, 1983; De Kleer 1986) can be assured. Truth maintenance guards the consistency of the model and is an important feature of user modeling attempts for specific domains as well as user modeling shells such as GUMS (Finin 1989) and TRUMP (Bonarini 1987).

6.2.4 Instantiating the user model: Collecting evidence in dialog
The context of user modeling in dialog behavior should be clarified before we can show how user models are instantiated, updated and exploited. Most of the systems that have been developed so far deal with user-system interactions in which the system is to assist the user in some way. For instance, the system provides information that the user asks for: *How can I remove a file?* is a possible query to the UNIX CONSULTANT (Chin 1989). Another communicative goal is explanation. The following is a sample question addressed to Quilici’s (1989) UNIX ADVISOR: *I tried to remove a file with the rm command. But the file was not removed and the error message was “permission denied”. I checked and I own the file. What's wrong?* To this, the system replies: *To remove a file, you need to be able to write into the directory containing it. You do not need to own the file.*

In the context of interchanges such as these, a system is supposed to draw a user profile. This includes a model of what the user knows; in the above example, the user knows which command to use to remove a file. It also includes what the user wants to achieve (e.g. the user wants to remove a file) and how he plans to achieve that goal (e.g. by using the rm command). At the same time, the system
has to infer, among other things, what knowledge is lacking and hence has to be provided to the user, and also what the user’s misconceptions are that need to be corrected by means of an explanation; in the above example, the user misconceives the relation between deleting a file and owning it. A similar interaction arises in a natural language interface to a database: Which students got a grade of F in CS105 in Spring 1980? is a possible question envisaged by Co-op (Kaplan 1983), to which the system may respond: CS105 was not given in Spring 1980 rather than None, so as to address the user’s misconception of the course time, and correct the premises of the question.

In the present case, the system faces the non-trivial task of determining why the user holds a particular belief or assumption, by trying to explain why the system itself does not hold that assumption (Quilici 1989). Hence, this example exemplifies a system that does not only construct a stereotype of the user, but also infers the causes of a user’s beliefs, for which the system has to go beyond a static stereotype in order to determine discrepancies between its own beliefs and assumptions and those apparently held by the user. Similar systems have been devised to detect users’ object-related misconceptions. ROMPER (McCoy 1989), for example, tries to explain to the user why his belief that a whale is a fish is actually wrong. Attention has also been paid to planning related misconceptions (Pollack 1986; Wilensky 1983; Wilensky et al. 1984; Chin 1989). The latter work stresses once again the importance of planning and plan recognition in natural language dialog systems (Carberry 1983, 1988, 1990; Grosz 1977; Robinson 1981; Allen 1983; Allen & Perrault 1980; Sidner 1983, 1985; Litman 1986; Allen & Litman 1986).

6.2.5 User modeling in simulated dialog systems

In recent years, the notion of adaptivity has become more important when building spoken interactive systems. Online, dynamic adaptation can be realised in the system’s ability to classify users into appropriate categories, for instance, on the basis of their navigation choices or a list of specific keywords, so as to produce personalised access to information sources and to enable filtering and recommendations within web browsing (Moukas & Maes 1998). Recommendation systems (e.g. Dai & Cohen 2003) track preferences of a group by comparing the selected items of one user to similar items selected by the other users.

In the EU-project DUMAS (Jokinen & Gambäck 2004) the main goal was to develop a prototype interactive email system, ATHOSMAIL, with components that
would make the user's interaction with the system more flexible and natural. The purpose of the User Model component in ATHOSMAIL (Jokinen & Kanto 2004) is to provide flexibility and variation in the system utterances, to allow the users to interact with the system in a more natural way, and to allow developers to implement and test machine learning techniques.

Various kinds of learning experiments have been conducted within dialog research, especially on optimizing dialog management strategies such as confirming or providing helpful information, by dialog simulation using reinforcement learning (e.g. Levin, Pieraccini & Eckert 2000; Scheffler & Young 2002; Litman et al. 2000; Walker, Fromer & Narayanan 1998; Williams & Young 2007). In this context user modeling usually refers to the building of a component that simulates the user's behavior, in order for the system developers to quickly test and compare various dialog strategies without engaging real users. Such simulated user models are particularly studied in reinforcement learning, since the algorithm requires hundreds of test cycles to converge to a solution, and it is difficult and time-consuming to get enough examples with human users.

6.3 Multimodality and affect

Communication includes more than mere verbal language use, and it is important to study such aspects of communication that are not only propositional but conveyed by multimodal means like facial expressions, head movement, hand gesturing and body posture (as argued by Allwood et al. 2007). Non-verbal communication, such as hand, head, and body gesturing or prosodic variance in speech, expresses the speaker's emotions, attitudes, and mental states, and also signals the speakers' social relations and interaction strategies with respect to other partners. Although non-verbal signaling can be conscious action, the speakers are not always very aware, and sometimes not even in control, of these signals. For instance, blushing or blinking of the eyes can be uncontrolled reactions conveying embarrassment.

Besides propositional language use in interactive situations, multimodality emphasizes the importance of prosodic and paralinguistic information in the processing of human conversational interactions, above and beyond the propositional information contained in the speech. Facial expressions of emotion and eyes in particular are central non-verbal communication channels in social interaction, also
including paralinguistic aspects of communication such as prosody and various vocalizations (laughs, sounds without propositional content) as well as visual cues. The use of bodily gestures and non-speech vocalisations can combine to show a speaker's attentive states and discourse intentions in a way that can be processed alongside the 'text' of the dialog to provide a rich source of interaction information. Research on multimodal dialogs and verbal and non-verbal signals that regulate the flow of information, is also crucial. The extra-propositional content in a spoken dialog is an essential component for smooth turn-management, of the shared context, and building of rapport and social bonds.

The use of prosodic and syntactic features in turn-taking and backchanneling has been studied from early on (e.g. Koiso et al. 1998), as well as establishing mutual eye-contact to indicate the end of turn and the speaker's intention to take turn (Kendon 1967). The importance of gaze in establishing the focus of shared attention in communication is learnt in infancy: gaze direction serves to frame the interaction and establish who is going to speak to whom (Trevarthen 1984). Jokinen et al. (in print) noticed that gaze helps to predict turn-taking, and is relevant especially in order to distinguish the speakers' normal hesitations from their willingness to give their turn to the conversational partner. Thus, eye-gaze may provide rich articulated feedback (Goodwin 2000) and support conversational and emotional signals conveyed via eyebrows (Ekman 2003).

The use of gestures in interaction is also well studied (Kendon 2004; Bavelas 2000). The meanings and functions typically assigned to pointing gestures concern referential pointing, giving directions, telling off someone, and commanding to pay attention. However, pointing gestures also seem to have an important function in communication management, i.e. on a conversational metalevel (cf. Kendon 2004).

Affect refers to a feeling or emotion, the way we perceive the world and form judgments, and it has a major impact on how well we are able to perform tasks: negative affect can make simple tasks difficult, positive affect can make difficult task easier. It is a well-known fact that humans assess interactions not only on the basis of efficient task completion but on the positive feeling and pleasant atmosphere during the conversation. What we remember from the interactions is not the length of conversation or the number of speech acts, but rather, how smoothly the communication took place so as to help us to reach the original task goal: if the partner does not understand what we are saying, or provides some unintelligible mumbling as a response, or implies that our problem is very silly, then the interaction may
not be rated highly, even though we would eventually get the task done. Positive affective systems also seem to change cognitive parameters of problem solving so that they emphasize breadth-first thinking and examination of multiple alternatives. This means that more complex tasks that require interactive systems can also be easier if the interaction itself supports a positive approach.

So far much of the interactive system design and development has focused on task completion, and the fluency of communication has been attached to producing complete and non-hesitational spoken utterances rather than recognizing multimodal signals. Typical dialog systems tend to make rather limited use of nonverbal and circumstantial information, but this is understandable as they operate on task-based scenarios, such as queries on bus timetables, banking, or airline booking, in which the interaction protocol can be fairly accurately designed in advance, and the user’s main goal is to get the task done.

However, if the task becomes more complicated and includes planning, or the goal of the interaction is to chat rather than accomplish a task, the interaction models need to take into account the situation in which the interaction takes place. It has also been argued that the users easily anthropomorphize the computer, i.e. assign human characteristics to it (Reeves & Nass 1996). When using natural language and speech, requirements of the interactive capabilities of such an automatic system may increase.

From a computational point of view, contextual information and situated language use have also become more important in the course of technology and application development. The ubiquitous computing paradigm assumes that the future environments will contain several context-aware devices that communicate with each other and with users (Weiser 1991). Interaction management thus requires that multimodal aspects of communication are taken into consideration, and that the user’s intentions, conversational strategies and involvement in the interaction can be detected and appropriately interpreted. The context of the interaction becomes more complex as the environment can dynamically change, which is challenging for the modeling.

Besides the embedded context-sensitive applications in the natural environment, also embodied conversational agents and interactive robots have become a popular application area. They try to recognize and respond automatically to the user’s affective states, thereby improving the quality of the interaction.
Conversational agents incorporate facial expressions (Andre & Pelachaud 2010; Gustafsson et al. 1999), but other multimodal aspects are also taken into account. Research and experimenting with the situated agents include automated understanding of referential utterances as well as gaze and gesture-based human-robot interaction (see e.g. Wilcock and Jokinen, submitted).

Another example of virtual agents is the sensitive artificial listener system SEMAINE-3.1, developed in the European FP7 SEMAINE project (Schröder et al. 2011). This system focuses on dialogs where the system needs to be able to interpret and produce appropriate emotional and non-verbal behaviors in order to maintain conversational dialogs. The system detects the user’s emotions and non-verbal signals from voice, facial expressions, and head movements, and the embodied conversational agent replies with expressive synthetic voice accompanied by suitable facial expression. While the user is talking, the agent also produces audio-visual feedback and emotionally expressive utterances so as to encourage the user to talk more (Scherer 2005).

7. Discussion and conclusion

The development of intelligent interactive systems can be traced as the development of computational pragmatics. The more elaborated systems are designed, the richer models of language interaction are needed: the context in which the interaction takes place and the various modalities to convey the message to the partner must increasingly be taken into account. In this chapter we provided an overview of the development of computational pragmatics by studying the different types of interactive systems and their requirements for suitable pragmatic models. We highlighted the differences by presenting various techniques and models for enabling interaction and discussed their applicability and use in developing different applications.

We have focused our discussion of computational pragmatics on dialog systems because most research seems to take place in this context. This does not deny the usefulness of computational pragmatics in other settings, such as machine translation. Although much recent work in machine translation uses shallow statistical approaches, it is obvious that deeper methods aimed at quality translation will
need to rely on inferences using pragmatic knowledge to identify corresponding appropriate speech act realizations in source and target languages.

The more complex systems are constructed, the more complicated it becomes to properly evaluate systems and try to determine appropriate features that guide user assessments. Especially for systems that exploit natural language understanding and reasoning, multimodality and complex task domains, evaluation is not a straightforward matter of getting a task done, but also involves the user's experience of the system and interaction itself. The user's perception of the system depends on the system's communicative capabilities related to the underlying task and which can vary from quick and simple prompts to natural intuitive interaction.

Walker et al. (1997) reviews proposals for dialog evaluation that can be objective or subjective. Objective evaluation metrics can operate without recourse to human judgements and are usually based on logs by the system itself, such as response times, turn-taking, failures and error messages, and matches of answers to reference answers. Subjective metrics are primarily based on judgements collected from users, for instance by means of evaluation forms, after they have used the system for particular task. A weighted function takes into account task-based success as well as multiple cost measures to assess the system's functioning and suitability. The standard criteria are usefulness (what the system is for), efficiency (how well the system performs), and usability (user satisfaction). Measures of dialog quality and dialog efficiency best predict a user's overall satisfaction (Litman et al. 2002). However, users seem to tolerate difficulties such as long waiting times and even mere errors, if the system is interesting and the users are motivated to use it. The users thus also assess the system's quality in respect to the service that the system provides: they look at the system from the point of view of its practicality and usefulness in helping them to achieve certain goals or gain some benefit, even if the application is not optimal.

Computational pragmatics continues to be a field rich with research challenges. More comprehensive models will rely on the collection and annotation of more data. Building large databases and sharing these with the community are clear desiderata. Progress will also depend on novel techniques and methods to create richer models for dialog management including non-verbal communication.
Note

1. The present chapter is a considerable reworking of Gillis, Daelemans and De Smedt (2009). The authors thank Steven Gillis and Walter Daelemans for the reuse of materials from the earlier earlier chapter.

References


