Natural Language and Dialogue Interfaces

Kristiina Jokinen

University of Helsinki

Department of General Linguistics

PO BOX 9

FIN-00014 Helsinki

Finland

Tel.: +358-50-3312521

Fax: +358-9-19129307

Email: kristiina.jokinen@helsinki.fi

and

University of Tampere

Department of Computer Sciences, TAUCHI

Finland

Abstract:

The design and development of natural interactive systems requires that the specific aspects of natural communication are taken into account. Besides the capability to understand and generate linguistic expressions, natural language use includes cooperation and planning of complex actions on the basis of observations of the communicative context, i.e., communicative competence. This Chapter discusses natural language dialogue interfaces and develops the view of interactive systems as communicating agents which can cooperate with the user on a shared task. Natural interaction is considered an approach to interface design which attempts to empower different users in various everyday situations to exploit the strategies they have learnt in human-human communication, with an ultimate aim of constructing intelligent and intuitive interfaces that are aware of the context and the user's individual needs. The notion of natural interaction thus refers to the spoken dialogue system's ability to support functionality that the user finds intuitive and easy, i.e., the interface should afford natural interaction. The view is supported by an evaluation study concerning a multimodal route navigation system.

Keywords:

intelligent interactive systems, natural language communication, human-computer interfaces, speech, multimodality

1. Introduction

Dialogue management technology has already several years enjoyed such a level of maturity where speech-based interactive systems are technologically possible and spoken dialogue systems have become commercially viable. Various systems that use speech interfaces exist and range from call routing (Chu-Carroll and Carpenter 1999) to information providing systems (Aust et al. 1995; Raux et al. 2005; Sadek 2005; Zue 1997) and speech translation (Wahlster 2000), not to mention various VoiceXML-type applications which enable speech interfaces on the web (VoiceXML Forum[1]). The common technology is based on recognizing keywords in the user utterance and then linking these to appropriate user goals and further to system actions. This allows regulated interaction on a particular topic, using a limited set of words and utterance types (see also Chapter X "Speech Input to Support Universal Access" of this Handbook). While spoken dialogue systems deploy flexibility and modularity of agent-based architectures (such as DARPA Communicator, Rudnicky et al. 1999; Seneff, Lau and Polifroni 1999), the applications are usually designed to follow a step-wise interaction script and to direct the user to produce utterances that fit into predefined utterance types, by designing clear and unambiguous system prompts. Spoken dialogue technology also uses statistical classifiers that are trained to classify the user's natural language utterances into a set of predefined classes that concern possible topics or problem classes that the user may want to get help about. In particular, they are deployed in the so called How May I Help You (HMIHY) technology (Gorin, Riccardi and Wright 1997), which also includes context-free grammars to guarantee high degree of accuracy of speech recognition and word-spotting techniques to pick the requested concepts from the user input. The HMIHY-type interfaces have been influential in the development of speech-based interactive

[1] VoiceXML Forum http://www.voicexml.org/

systems, although typically they do not include deep language understanding components.

In technology and commercial application development, the focus has been on relatively simple tasks that would not need sophisticated interaction capabilities beyond individual word and phrase recognition. Some evidence has been found for supporting the view that the users actually prefer a command-based interface to one with human-like communication capabilities: the users want to get the task done, efficiently and with as little trouble as possible, and it does not matter whether they can express themselves using natural language with the same freedom and fluency as in human-human communication. It must be noted, however, that in similar situations with a human operator, the users do not expect fancy conversation capabilities either but decent task completion: the users would reward quick and accurate service and not necessarily chatty but socially nice conversations. In research communities, on the other hand, interaction models strive for more human-like spoken language capabilities, to understand spontaneous speech, and to model reasoning processes. The focus has especially been on error management and handling misconceptions, on conversational features of speech as well as on multimodal aspects of communication. This research helps in understanding human communicative behaviour, but it also serves as an inspiration for advanced spoken language technology that aims at developing adaptive interfaces and handling spontaneous natural language utterances.

However, both practical and research prototypes often fail to reach the level of user satisfaction that would allow the users to enjoy the interaction. The reasons have been pinned down to bad speech recognizer performance, as well as to straightforward dialogues which are directed towards task completion. In system initiative dialogues, the users are also required to follow the system (designers) conceptual models, instead of being able to express their goals in their own way. It is thus assumed that solutions, where natural language communication is supported both on the technical and conceptual level, are likely to succeed better. There are several reasons for this. First, speech creates an illusion of real interaction capabilities, and the users apply their intuitive human language strategies to interact with the system, even though they may be well aware of their partner to be a computer (cf. studies by Reeves and Nass 1999). Although the users can learn to speak to the computer (i.e., learn the "computerese" language), and may prefer a straightforward interface that is not too human-like, it is the human manner of interaction that is used as the standard for comparing the system performance; in other words, the view point for evaluating human-computer interaction is the human interaction. Second, applications in which spoken interaction would show its power are characterized by complex domains which are difficult to model in the detailed and exhaustive way required by the state-based technology. For instance, trip planning and assistance in machine maintenance fall in this category: the speakers may not know what are the different options and alternatives concerning their preferences for a planned trip, or how to classify maintenance problems using technical terminology, but can reach satisfactory solutions in these tasks through natural language communication. However, planning requires separate knowledge of the possible plans and their prerequisites, and a number of technical maintenance operations require specialized knowledge of the relevant terminology and functioning of the machine. The HMIHY-type modelling of interaction by combining domain information and dialogue acts in few dialogue states quickly becomes infeasible: the number of possible dialogue states is huge and the listing of all possible paths through the state space, i.e. all possible ways to conduct a dialogue with the user, simply becomes impossible. In these applications, separate models for reasoning about dialogue strategies and task planning would be necessary, and a natural language interface would allow flexibility both for the user to express requests and statements, and for the system to present complex information. Third, in interface design, a new metaphor has also appeared to replace the traditional “computer as a tool” -metaphor, namely that of “computer as an agent” (Jokinen 2003, 2008a). The new metaphor regards the human-computer interaction as a cooperative activity between two agents: the system and the user. The system mediates between the user and the backend application, and assists the users in achieving their goals. The system is expected to act in a manner of an intelligent rational agent, so the metaphor presupposes the system's capability to communicate through natural language.

The challenge that speech and language technology faces in this context is not so much in producing tools and systems that enable interaction in the first place, but to design and build systems that allow interaction in a natural way: to provide models and concepts that enable experimentation with complex natural language interactive systems, and to test hypothesis for flexible human-computer interaction. The non-natural aspects of interfaces are usually traced back to the lack of robustness in speech recognition, deficient natural language understanding, and inflexible interaction strategies, but there is still a long way to go before interactive applications can pass their human counterparts in robustness and versatile performance. In addition to improved interaction strategies, natural language interfaces are also required to be extended in their knowledge management and reasoning capabilities, so as to support inferences concerning the user's intentions and beliefs behind the observed utterances. Since interaction can take place via speech, text, and signing, special attention is also to be paid to multimodality (see Chapter “Contributions of “Ambient” Multimodality to Universal Access” for more details about multimodal user interfaces).

The design and development of more natural interactive systems requires that the aspects of natural communication that facilitate intelligent and intuitive interaction are taken into account.

The goal of building natural interactive systems thus comes close to studying intelligent interaction in general. This demands better understanding of how people ground their intentions in the environment they live in, and how they communicate their intentions to other agents, which may well be automatic devices and interactive systems besides human beings. In the ubiquitous technological context (Weiser 1991), it may seem less bizarre to think that the design of natural language interactive systems is, in fact, analogous to the construction of intelligent systems in the artificial intelligence research. The use of natural language consists of planning and performing complex actions on the basis of observations of the communicative context, and this accommodates with the notion of intelligent agents in artificial intelligence as defined in the authoritative text book (Russel and Norvig 2003:vii): “The main unifying theme is the idea of an intelligent agent. We define AI as the study of agents that receive percepts from the environment and perform actions. Each such agent implements a function that maps percept sequences to actions, --”. Interactive natural language systems can basically be defined in a similar way as implementing a function that maps the agent's understanding of the communicative context to communicative actions.

Without going deeper into the philosophical aspects of intelligent interaction with machines, it should be pointed out that naturalness is a problematic concept even when consideration is limited to the techniques and models that enable implementation of intelligent interaction. Often, it simply refers to the use of natural language as a mode of interaction, but can also be understood in a wider sense that focuses on the natural and intuitive aspects of the interaction in general. It can be attached to modelling human-human communication and building systems that try to mimic human communication, and it can also refer to applications that try to take advantage of the users’ natural ways of giving and receiving information. The wider perspective of natural interaction, however, requires investigations on what it actually means to communicate in natural ways in different situations. For instance, natural language is by far the best means of explaining and negotiating complex abstract ideas, while gesturing is more natural when pointing and giving spatial information. On the other hand, spoken language is not suitable if one is concerned about disturbing others or privacy issues, and much of the speakers' attitudes and emotions are also expressed by facial expressions and body posture, which are natural means for conveying non-verbal information. In human-computer interaction, quite the opposite is the case: graphical interfaces with touchscreen, mouse and keyboard are more natural than the abstract ways of expressing oneself in natural language: they seem to provide the user with more concrete control over the task than language-based conversations. When considering universal access to digital information in general, the natural means of human-computer interaction include possibilities for interacting with both spoken and written text, as well as with various other interactive modes such as tactile or gaze-based interaction. Novel interface types are not only introduced as alternatives for users with disabilities, but as supporting natural interaction in general. For instance, Milekic (2002) talks in favour of tangible interfaces that allow grabbing, and emphasises that they are easier, more natural and more effective to use, since they apply biological knowledge of what it feels to touch and grab things: this enables manipulation of objects without explicit formalization of what one is doing.

In order to develop natural human-computer interfaces, it is thus necessary to consider the communicative situation as a whole. Jokinen (2008a) discusses the system's communicative competence which includes the following: physical feasibility of the interface, efficiency of reasoning components, natural language robustness, and conversational adequacy. The first aspect refers to the enablements for communication such as the user having access to digital information and being able to use the system in a natural way, while the second one refers to algorithms and architectures that enable the system to conduct robust and natural interaction with the user. The two last aspects take the system's language processing and interaction capabilities into account: analysis and generation of utterances, managing dialogue strategies, and adaptation to the user. According to this view, it is important to investigate the coordination of different input and output modes (speech, text, pen, touch, eye movement, etc.), so as to utilize appropriate modalities for natural exchange of information between different users in a wide variety of communicative situations. Consequently, natural interaction can be considered as an approach to interface design which attempts to empower different users in various everyday situations to exploit the strategies they have learnt in human-human communication, with an ultimate aim of constructing intelligent and intuitive interfaces that are aware of the context and the user's individual needs. The notion of natural interaction thus refers to the system's ability to support functionality that the user finds intuitive and easy, i.e., the interactive system should *afford* natural interaction, cf. Norman (2004). Affordance contains natural language as an essential mode for the user to interact with the system.

This Chapter discusses natural language dialogue interfaces in detail. The Chapter is structured as follows. Section 2 provides an overview of natural language interaction and natural interfaces. Section 3 develops the view of complex interactive systems as communicating agents, and it also provides a short overview of the Constructive Dialogue Model, and discusses contextual understanding of dialogues. Section 4 presents a route navigation system as an example of natural language applications, and discusses its evaluation from the point of view of the user's expectations and experience. Finally, Section 5 draws conclusions concerning system evaluation and the system’s communicative capability.

2. Natural Language Interaction

2.1 Linguistic approach

Natural language is both a cognitive and cultural phenomenon: it is purposeful activity by individual speakers, but it is also a speaker-independent entity regulated by the norms of the language community. It is used to communicate one’s ideas and thoughts, and it is also a means to build social reality through interaction. The focus of linguistic research has thus been on the structure of language on one hand, and on the use of language on the other hand. The former includes research on the rules and regularities of morphology (word formation) and of syntax (phrase and sentence parsing), while the latter concerns semantic and pragmatic inferencing as a function of the interaction and communicative context.

An important aspect of language use is to understand the propositional meaning that the observed linguistic entities (written sentences, spoken utterances, signed expressions) convey. The meaning can be constructed on the basis of the order of the elements (*John ate a fish* vs. *A fish ate John*) but is also dependent on the context in which the elements occur: e.g., the word *table* may have different interpretations in the request *You should move that table somewhere else* (a piece of furniture or a figure in a document), and the phrase *thank you* may have different functions (a sign of gratitude or a sign of wanting to finish a conversation). Hence, natural language communication does not only consist of attaching words together according to certain grammar rules, but of interpreting messages and relating their meaning to the existing world. It is an activity through which individual speakers create knowledge representations of the world. Linguistic meanings exhibit both individual and social dimensions: they are simultaneously represented in the cognitive processes of the speakers as well as in the cultural schemes that the speakers have created, and continuously create, through interaction and communication. On the other hand, in linguistic anthropology and conversation analysis, Gumperz and Hymes (1972) have suggested that language exists only in the interaction with other members of the language community. Linguistic meaning is the shared understanding that emerges among the speakers in the communicative situations, and thus it is learnt with respect to the speakers and situations. However, this view does not explain why language is also independent from the individuals and their interactions: it is not the case that meanings need to be created and learnt separately in each speech event, but the speakers can, in fact, rely on some shared code that already exists within the language community. In other words, there is a link between the speaker's individual knowledge of the world and the development of shared reality, culture and society, i.e., the natural language used in the language community.

In the classic work by Grice (1957, 1989), two types of meaning are postulated: the *linguistic meaning* refers to the semantic interpretation of the utterance, and the *speaker meaning* refers to the pragmatic interpretation whereby the intentions of the speaker are spelled out. Successful communication requires that the speaker has an intention to deliver a message, and that the listener recognizers the speaker's intention to communicate something. The signals sent by the speaker are thus meant to be understood as symbols with certain meanings, and the listener is expected to react to them in a relevant manner.

The intentions of the speaker are encoded in the discourse functions of the contributions. Following the Speech Act theory of Austin (1962) and Searle (1979), utterances function as actions which have certain prerequisites and which are aimed at fulfilling the speaker's intentions. Utterances may contain performative verbs which explicitly indicate the act performed (such as promise, baptize, etc.), but usually the act is to be conventionally inferred on the basis of the utterance content and its context (requests, statements, acknowledgements). In dialogue research, the term *dialogue act* is used to emphasize the fact that the original notion of speech acts is extended and modified to cover various dialogue properties (see, e.g., Bunt and Girard (2005) for a taxonomy and recognition of dialogue acts). Dialogue acts have been influential in the development of plan-based dialogue systems (e.g., TRAINS, Allen et al. 1995) and, although they may not constitute an explanation of the dialogue as such, they can be used as convenient labels and abstractions of the speaker's mental state in order to structure dialogues and learn dialogue strategies.

Natural language enables speakers to exchange meaningful information for the purpose of achieving certain (linguistic or non-linguistic) goals. Consequently, dialogue contributions are linked together: the speakers act in a rational and consistent manner in order to reach their goals. The dialogues show global coherence in the overall dialogue structure, and local coherence in the sequences of individual utterances (Grosz and Sidner 1986). The former captures a general task structure and is often modeled with the help of a task hierarchy, a dialogue topic, or a focus stack, while the latter realizes the speakers' attentional state, and builds up on the cohesive means between consecutive utterances like the use of pronouns (*A: Take the bus number 73. B: Does* **it** *stop by the hospital?*) or ellipsis (*A: yes, it stops by the hospital. B: and [does it stop] by the station?*).

The dialogue is also teamwork, and the partners collaborate on the underlying task (see articles by Cohen & Levesque and Grosz & Sidner in Cohen, Morgan and Pollack 1990, as well as recent work on the collaborative interface agent Collagen, Rich et al. 2000). The participants also cooperate on building shared knowledge via grounding and feedback (Clark and Schaefer 1989), as well as on co-producing dialogue contributions (Fais 1994), and being engaged in the communicative activity in the first place (Ideal Cooperation, see below Section 3.1).

Finally, language is dynamic, and its changing nature makes the modelling of natural language interaction difficult. The semantics of words evolves, new concepts develop, and also different dialogue strategies are learnt so as to cope with novel circumstances. It is impossible simply to itemize the facts necessary for modeling relevant actions and events, so learning and adaptation are necessary. As argued in Jokinen (2000), interactive systems need to be equipped with a dynamic update procedure that emulates learning through interaction. Learning to associate concepts with situations in which the concepts are used enables adaptation to different circumstances. Adaptation and learning are pertinent features especially for applications that aim at coping with complex tasks, such as negotiations and planning, but they can improve robustness even in practical information providing systems like train timetable or restaurant guides. For instance, a user who is engaged in a dialogue concerning good restaurants may suddenly ask about train timetables and the leaving of the last train. Usually, dialogue systems treat such questions as off-domain interruptions, and the user is requested either to continue on the original topic, or to confirm that she wants to quit the dialogue. However, the question may also be motivated by the user's need to catch the last train after an evening out, and it indicates indirectly the user's preferences for the location of the restaurant. Thus it may be linked together with the topic of restaurant choices as a possible world knowledge association for the user. These kinds of associations are of course impossible to anticipate or list exhaustively in advance, but the associations between the domain and world knowledge can be learnt from interactive situations. The bond between the two items can be activated and reinforced if the user happens to appear in a similar situation again, but if the co-occurrence appears to be an unexpected one-time event, the bond between the concept and its associated context is decreased. Similar kind of learning experiments have been conducted within dialogue research, especially on optimizing dialogue management strategies such as confirming or providing helpful information, by dialogue simulation using reinforcement learning (e.g., Levin, Pieraccini and Eckert 2000; Scheffler and Young 2002; Walker, Fromer and Narayanan 1998; Williams and Young 2006).

2.2 Natural language interfaces

The extensive use of natural language suggests that the role of a natural language front-end is to be redefined in interactive systems. It is not simply an interface that connects user commands to a set of possible automatic reactions, but a special software component that initiates, maintains, and records interaction between human users and computational applications. In other words, dialogue systems contain a particular dialogue management component, which initiates interaction and creates abstract representations that are further manipulated and processed in the reasoning components distributed in the system's architecture.

As a consequence, the interface has to be distinguished from the system's language capabilities. The interface refers to different physical devices through which human-computer interaction is enabled, such as the computer screen, mouse, keyboard, and microphones. Natural language, however, should not be mixed with the medium, since it brings in an extra dimension compared with the command and menu-based systems: that of symbolic communication (c.f., discussion on the medium, code, and modality in Jokinen 2008b). Language is a particular code for interaction, analogous to a set of gestures or colour signs, except that its expressive capacity is richer and more varied. Besides offering a wide range of conventional symbols for presenting information, a natural language interface also presupposes minimum conversational capability, i.e., an ability to interpret utterances within the dialogue context and to hypothesize possible intentions behind the utterance usage.

A top-level view of a dialogue system is presented in Figure 1. The natural language front-end consists of an input analyzer and an output generator. The former includes language understanding components such as a morphological and syntactic analyzer, and a topic spotting and semantic interpretation module, while the latter includes components dealing with the planning and generation of system utterances. Spoken dialogue systems also include a speech recognizer, which often integrates language understanding components, and a speech synthesizer which speaks out the system utterances. The dialogue manager is the component that manages interaction: it deals with decisions on how to react to the user input and how to consult the backend database, and it also updates the dialogue history. The task manger is a special reasoning component that takes care of the reasoning and access to the backend application. It may be part of the dialogue manager, especially if the application related reasoning is included in the dialogue strategy design, but to emphasize the need for intelligent reasoning and inference, it is depicted as a component of its own in Figure 1. Separate information storages are usually maintained to keep track of the task and dialogue-related information dynamically created and exchanged in the course of the dialogue, as well as to encode user preferences.

Insert Figure 1 about here

Figure 1 also depicts interface languages between the components. The user and the system interact with natural language, which is parsed and semantically analysed in the natural language front-end. Semantic interpretation is often a keyword or pattern-based "shallow" interpretation process, but can also deploy sophisticated parsing techniques to uncover syntactic-semantic relations between the constituents. Spoken language input would first go through a speech recognizer, which produces an n-best list of possible recognized utterances with a confidence score or a network of words from which the correct interpretation can be picked. The interpreted result, the semantic representation of the utterance, contains semantic predicates and arguments representing the utterance meaning as well as a dialogue act label representing the speaker's intention. For instance, if the user has informed the system *I'd like to go to the station*, the semantic representation may look as a conjunction of predicates as follows:

```
[inform(x,y), user(x) & system(y) & travel(t,x) & destination(t,s) & station(s) &
def(s)]
```

The predicates may also be linked to ontological concepts, in which case the representation should be called "conceptual representation". From a linguistic point of view, the two representations are at different descriptive levels, but in practical work, the fine distinction is usually not necessary.

The semantic (or conceptual) representation can further instantiate a Frame, i.e., a representation of the task or application knowledge that can be used as a basis for various reasoning processes. For instance, a travel-frame may contain slots for a departure and a destination place, a departure time and an arrival time, as well as for a route between the two places, and look like the following flat structure:

```
[depPlace=P; depTime=T; destPlace=D; destTime=S; route=R]
```

The semantic representation may also trigger a plan recognition process and instantiate a plan for the purpose of reaching a certain goal. A plan may consist of a sequence of actions to be pushed and popped in accordance with the plan execution, or be linked with a particular frame, so as to anchor the acts with a relevant conceptual framework. The dialogue manager may also maintain an Agenda, which is simply a list of actions that it needs to perform, including dialogue acts with the user, system acts to update the information storages, and database commands to access the backend database. The actual implementation of the different representations is commonly performed with the help of XML, which provides a standard interface language that is easy to process and integrate with other system modules such as the speech and multimodal components.

This Chapter does not go into details of the different dialogue management techniques (scripts, frames, and agent-based approaches), but the reader is referred to the overview in McTear (2004). As for references and implementation of dialogue systems, see also Jokinen et al. (2002), Jokinen (2003, 2008b).

3. Complex Systems as Communicating Agents

Interactive systems are mostly task-based applications which model straightforward question-answer interactions, and may thus seem rather simple from the point of view of human conversations. The HMIHY-type applications inherently expect quick and clear interaction, and the efficiency of the interface, i.e., the efficiency of the natural language dialogue, is measured in regard to the length of the dialogue. As mentioned above, expectations in these tasks concern efficient and clear interaction rather than socially adapted conversations, and even if the partner is a human service agent, efficiency counts as a major factor for the satisfactory service. It is thus plausible that the reason for natural language interfaces not being as popular as one might expect on the basis of the fundamental status of natural language in human communication is related to the fact that the underlying tasks are too simple and familiar to support spontaneous use of the whole range of human conversational capabilities. Moreover, human-computer situations are single and serial: the users interact with the system one at the time, and perform one task at the time. Natural language, however, is better suited for a different type of communicative behaviour: it shows its power in complicated multiparty multitask situations where various issues need to be negotiated and also problems can occur. Natural language is used to clarify misunderstandings and to correct wrong information, as well as to reason on the situation itself. It is important to notice that in order to solve problems, the partners should be able to discuss about the information that has been exchanged in the dialogue, and for this, it is essential to reason on what was said, why it was said, and what could have been said instead, i.e., to master meta-level communication that concerns the language and communication itself, not just objects and actions in the world. Moreover, natural language is essential for the coordination and communication of social interaction, to understand subtle signals related to acceptable and polite interaction, and to avoid imposing or embarrassing the partner.

Returning back to ubiquitous and context-aware communication mentioned earlier, natural language communication seems to provide a realistic interface for the intelligent environment. The ubiquitous computing paradigm envisages that pervasive and mobile communication technologies will be deployed through a digital environment which is aware of the presence of mobile and wireless appliances (see also Chapter X "Ambient Intelligence"). The environment is adaptive and responsive to the users' needs, habits and emotions, and ubiquitously accessible via natural interaction. The use of text, speech, graphics, touch, and gaze allows natural input/output modalities for the user to interact with the back-end application, and the users can choose the modality that best suits to the particular circumstance and to the users' preferences. Current research projects concern, e.g., how to make communication with the house a reality, while chatbots and conversational avatars have been introduced as future web surfing method. Many initiatives focus especially on mobile, ubiquitous, and location-aware services, enabling "invisible intelligence" in systems and applications of everyday life (such as cars, home appliances, factory automation, mobile phones).

The ubiquitous computing paradigm changes human-human communication, too. Novel aspects for interaction are brought forward and, consequently, the extent and type of social interaction is also about to change. The users can share their own (digital) data among friends and colleagues, and learn from the other members of the community by navigation, intelligent browsing, and direct interaction. Virtual communities are created, where interaction is rapid although not necessarily face to face, and one's identity may also be hidden behind different roles. This kind of on-line communication presupposes off-line processing of vast amounts of digital data, consisting of such data types as texts, music, photos, videos, and emails. The organisation of data should be automatic and fast, and allow human intervention in directing and guiding the processes according to individual preferences and needs. Interaction management with the application could thus support off-line organisation of the data and its retrieval according to some topical principles which relate to the conversational topic that the speakers are talking about. Below two relevant aspects of conversation in this respect are discussed, cooperation and context management, both of which affect the smoothness of communication in ubiquitous contexts.

3.1 Constructive Dialogue management

As already mentioned, language is purposeful behaviour by rational agents. For instance, Allwood (1976) defines communication as activity by rational agents bound by social obligations. A dialogue system's desired behaviour can also be grounded on the notion of cooperative and appropriate communication (Allwood, Traum and Jokinen 2000; Jokinen 1996, 2008a). Rational cooperation can be seen as emerging from the partners' communicative capabilities that maintain interaction on the basis of relevant and truthful information. It is based on the speakers' observations about the world and on their reasoning, within the dialogue context, about the new information being exchanged in the dialogue contributions. In human-computer interaction, cooperation manifests itself in the system properties that enable the user to interact with the system: robustness of data processing and appropriate presentation of the information to the user (Jokinen 2008a). Robustness is thus not only a quantitative measure of the system's response capabilities, but it also subsumes qualitative evaluation of the system's communicative competence.

In human-human conversations, cooperation refers to a participant’s overt behaviour that seems to convey the participant’s willingness, benevolence, and ability to provide relevant responses that address what the partner has questioned. According to Allwood (1976), the speakers are engaged in Ideal Cooperation if they:

(a) have the same goal,

(b) consider each other cognitively,

(c) consider each other ethically, and

(d) have trust that the partner behaves according to (a)-(d).

The first requirement means that the agents must cooperate *on* something, i.e., they must share intention to achieve a certain goal. Cognitive consideration refers to the agent’s deliberation on the fulfilment of the goal in the most reasonable way. An important dimension in Ideal Cooperation is ethical consideration, which obliges the agents to treat their partners as rational motivated agents as well. This means that rational agents should not only attempt to fulfil their own goals, but they should not prevent other agents from fulfilling their goals either. Besides bringing aspects of politeness, indirectness, voluntary help, etc., into rational acting, ethical consideration also accounts for the agents' seemingly irrational behaviour, such as volunteering for a tedious job to save someone else from doing it (increase own pain instead of pleasure), or choosing an inefficient method which would allow easier interaction with the others (act in an incompetent way). Ethical consideration thus functions as a counter-force to cognitive reasoning. Finally, the mutual trust binds the agents' acts together and provides a basis for understanding the speaker's meanings: by assuming that the partner communicates according to the Ideal

Cooperation, the agent can recognize intentions behind the communicative acts, deliberate on the possible alternatives, and decide on the appropriate response.

In interactive systems, cooperation and communicative obligations are usually hardcoded in the control structure, and the system's reactions are aimed at producing the most straightforward response. The system cannot reason on the meta-level about different cooperation strategies which, as mentioned above, would be necessary in negotiations and resolving misunderstandings and other problematic situations. In the context of mobile information technology, models of rationality and cooperation can improve the quality of practical systems: by implementing these principles in the reasoning process, interactive systems can be made more flexible and reactive. Constructive Dialogue Modelling (Jokinen 2008a) implements Ideal Cooperation as part of the construction of a shared model of two communication-related tasks: evaluating the partner's goal, and planning an appropriate response to it. The former concerns how the partner's goal can be accommodated in a given context, and results in strategic decisions about appropriate next goals. The latter concerns how the goal can be realized in different contexts. From the agent's point of view, communication consists of an analysis of the partner's contribution, the evaluation of the new information in regard to the agent's own knowledge and intentions, and reporting the evaluation results back to the partner in the agent's (communicative) reaction. While the evaluation of new information is motivated by the agent's cognitive consideration within the changed context, the reporting of the result is influenced by ethical considerations: the agent needs to inform the partner of the new situation, so as to allow the partner to work towards their shared goal with all the relevant information.

The agents respond to the changes in the context in which the interaction takes place, but they are also capable of planning and taking initiatives to fulfil their own goals. Since the evaluation of the exchanged information takes place in the context of the agent’s plans and goals, the agent's reaction may not always be as expected by the partner. In this case, Ideal Cooperation obliges the agent to provide a reason for the failure to act according to the cooperation expectations, possibly together with an apology, so as to allow the partner to re-evaluate the feasibility of her original plan. The re-evaluation may then lead to further negotiations in order to reconcile the participants' contradictory intentions. It is important to notice that mere refusal to act without any apparent reason would usually be interpreted as a sign of unwillingness to cooperate, and would result in an open conflict.

The main features of the CDM agents are:

- The agents are rational and cooperative,
- The agents exchange new information,
- Mutual knowledge is constructed through interaction,
- Utterances are locally planned and realized,
- The agents use general conversational principles (Ideal Cooperation).

In an ubiquitous context, the cooperation between human and the intelligent environment can be modeled based on the same principles of Ideal Cooperation and the CDM agents. The agents' actions in interactive situations are usually on-line reactions to contextual changes, i.e., to the new information that is being exchanged. For instance, a request to switch on the light or a question about the last bus is usually responded by switching the light on or providing the relevant information. In case the agent wishes to refuse the request, it is necessary to produce a relevant reason (e.g. the fuse is blown/Sorry, I haven’t got the timetables/I haven’t got a permission to give the information/I don’t think taking the bus is a good idea/It would be more economical and ecological not to switch on the lights, etc.).

3.2 Context understanding

Human-human communication involves smooth coordination of a number of knowledge sources: characteristics of the speakers, topic and focus of the conversation, meaning and frequency of lexical items, communicative context, physical environment, world knowledge, etc. The following human-human dialogue between a service agent and a customer (Interact corpus, Jokinen et al. 2002) exemplifies these aspects: the overall dialogue structure is non-deterministic, and the agent's guidance shows flexible and considerate interaction strategy.

A: *I'd like to ask about bus connection to Malmi hospital from Herttoniemi metro station – so is there any possibility there to get a bus?*

L: *Well, there's no direct connection – there's the number 79 that goes to Malmi but it doesn't go to the hospital, it goes to Malmi station*

A: *Malmi station? oh yes – we've tried that last time and it was awfully difficult*

L: *well, how about taking the metro and changing at Sörnäinen, or Hakaniemi if that's a more familiar place*

A: *Well Hakaniemi is more familiar yes*

L: *Ok, from there you can take the bus 73*

A: *73?*

L: *Yes it leaves Hakaniemi just there where you exit from the metro to the bus stops, next to the market place*

A: *So it's by the market place that 73 leaves from?*

L: *Yes*

A: *And it's not there where the other buses leave from in front of Metallitalo?*

L: *No, it's there right when you come out from the metro*

A: *And it goes to the hospital?*

L: *yes, it has a stop just by the hospital*

A: *Ok, it must be a better alternative than the bus we took to the station, we didn't know which way to continue and nobody knew anything and we had to take the taxi…*

L: *what a pity – there would have been the number 69 though. It leaves close to the terminal stop of number 79 and goes to the Malmi hospital.*

A: *I see, so 79 to the station and then 69?*

L: *yes*

A: *Are they on the same stop?*

L: *well not on the same stop but very close to each other anyway*

A: *close to each other? Ok, well thank you for your help.*

L: *thank you, goodbye*

A: *goodbye*

This dialogue shows how the speakers cooperate with each other on building a shared understanding of what is the best bus route, and also how they pay attention to the partner's emotional state and needs. For instance, the agent senses the customer's frustration and introduces a simpler route via Hakaniemi using metro and bus, but as the customer returns to her earlier frustrating experience, the agent provides information of this option, too. Language is also related to the context in which the dialogue takes place. The speakers make frequent references to the physical environment (*change in X*, *close to each other*, *Hakaniemi*, *Malmi station*), and the spatial and visual environment directs their interpretation and generation of linguistic utterances (*it's there right when you come out from the metro*). In other words, language is grounded in the communicative context. Grounding is a part of natural communication, exemplified by frequent situational and contextual references, and the whole range of different modalities used in processing and manipulating information (gestures, pointing, nodding, gazing, etc.). As mentioned above, in linguistics grounding also refers to building of the shared understanding of the dialogue goal (Clark and Schaefer 1989), i.e., to the agent's giving feedback of their understanding of the presented new information (*ok, I see*).

In interactive systems, the context is included in the general usage scenario. The assumed context may not be the one the user is in, however. Especially in mobile situations, the systems should interact by making dynamic references to the physical environment, and also allow pointing, gestures and gazing as input modes, since purely verbal expressions may become rather clumsy. Moreover, information should be presented in accordance with the communicative context, since the user’s attention is not necessarily directed towards the device and the service that the device provides, but is often divided between the service and some primary task such as meeting people. This requires awareness of the context and the user’s intentions, as well as an ability to interpret the user's verbal utterances with respect to the possible goals as are apparent in the context.

4. Example of an interactive navigation system

This section briefly presents one location-based service, the multimodal map navigation system MUMS, and its evaluation concerning cooperation in mobile context. The MUMS system (Hurtig and Jokinen 2006; Jokinen and Hurtig 2006) is a multimodal route navigation system which aims at providing the user with real-time travel information and mobile navigation in Helsinki. It was developed in a technological cooperation project among Finnish universities, supported by the Finnish Technology Agency TEKES and several IT-companies. The main goal was to build a robust practical application that would allow the users to use both spoken language commands and pen-pointing gestures as input modalities, and also output information in both speech and graphical modes. The system is based on the Interact-system (Jokinen et al. 2002) which aimed at studying methods and techniques for modeling rich natural language based interaction in situations where the interaction had not been functional or robust enough.

Insert Figure 2 about here

Figure 2 presents the general architecture of the MUMS system. The PDA-client only has a light-weight speech synthesizer, while the server handles all processing of the information. The touch-screen map can be scrolled and zoomed, and the inputs are recorded simultaneously and time stamped for later processing. The architecture is described in more detail in (Hurtig and Jokinen 2006). A screenshot of the system output is presented in Figure 3. It corresponds to a situation where the system answers a user question in spoken and graphical form: *The tram 7 leaves from the Opera stop at 13:46. There are no changes. The arrival time at the Railway Station is at 13:56*.

Insert Figure 3 about here

The evaluation of the system aimed at establishing the difference between the users' expectations and experience with the system, as well as studying how the background information affects the users' views about the system properties, especially its cooperation with the user. The users had varying levels of experience with speech and tactile interfaces, but none had used the MUMS or a similar system before. They were given a short demonstration of the system use before the actual evaluation sessions, and were divided into two groups depending on the background information given to them about the system they were about to use: one group was instructed to interact with a speech interface that also had a tactile input option, while the other group was instructed to interact with a tactile system with spoken dialogue capabilities. Both groups were then given the same scenario-based tasks (Kanto et al. 2003) that dealt with using the system to find means of public transportation to reach certain places. The expected and experienced system performance was measured by asking the users to fill in the same evaluation form twice: first to estimate their expectations of the system performance before the actual tasks, and again after completing the tasks. The questions were organized into six groups concerning the user's perception of the system's speech and graphical interface, the user's perception of the system's functionality and consistence, the user's perception of the system's responses and their appropriateness, the user's perception of system taking the user into account and the easiness of completing the task, the user's eagerness in future use of the system, and the overall assessment of the system.

The evaluation results are reported in detail in (Jokinen and Hurtig 2006), and show that user expectations were fulfilled and the system seems to have offered a genuinely positive experience. Here the differences in the perceived performance measures and the overall change between the user expectations and the actual use of the system are briefly discussed.

As expected, the users' prior knowledge of the system influenced their evaluations. Both groups gave very positive reviews for the system's multimodal aspects, and felt that both touch and speech are important in the interaction. The speech group, however, seemed to emphasize the contribution of the system's speech and graphical representation to the intelligibility of the system's output, and they were willing to use a tactile interface unimodally, i.e., they believed that a tactile interface was more usable even if not combined with speech. The tactile group, on the other hand, seemed to think that several modalities make interaction flexible and the system easy to use, but they also believed that a unimodal tactile interface would not be as usable as the multimodal one. There was also some evidence that the tactile group was more willing to use a unimodal speech system than the speech group, but the differences were not significant. The speech group also felt, more than the tactile group, that the system was slow, even though the response time of the system is not affected by the form of user input. This supports the claim that speech presupposes rapid and understandable communication. Moreover, it is interesting that the tactile group was slightly more positive at the use of speech input and output than what they had expected, whereas the speech group was disappointed with the use of the unimodal speech system. The speed of the system is thus important, especially if the users expect that the system is meant for spoken interaction. Analogously, the tactile group was more critical towards the map qualities, and in fact, the difference between the user's expectations and perceived system qualities was in absolute terms negative. The speech group, on the other hand, perceived the use of the map quite positively.

The results show that the priming effect comes to play a role in system evaluation. The tasks given to the two groups and the system that the groups used were exactly the same, but the expectations and the experience varied significantly between the groups. In general, system properties like understandability and pleasantness of the speech output are highly evaluated, and the users were unanimous that the system with both speech and tactile information is preferable to a unimodal one. However, compared with the speech group, the tactile group was positively surprised at the system capabilities and the system's functionality: the system was reported to be helpful, considerate, and cooperative, functioning in a consistent way and capable of recognizing the spoken input usually at the first try. In the conducted evaluations, speech recognition worked in a similar fashion for both groups, so the differences cannot be associated solely to speech recognition problems. Rather, the answer seems to lie in the predisposition of the users and their expectations concerning whether they interact with a speech interface that uses tactile input, or with a tactile interface that can speak as well. As mentioned above, the users automatically adjust their assessment of the system communicative capabilities with respect to what can be expected from a system, and the use of spoken language seems to bring in tacit assumptions of fluent human communication and expectations of the system possessing similar communication capabilities.

5. Conclusions

This Chapter has discussed natural language communication as an important aspect of an interactive system's functionality and usability.

When dealing with present-day interactive services, the users are often forced to adapt their human communication methods to the requirements of the technology. Simultaneously, with the development of ubiquitous environment, a growing number of interactive applications will appear, and expectations for fluent and intelligent communication become higher. It is thus necessary to support the system's communicative capability to understand and provide natural language expressions, recognise new and old information, reason about the topic of interaction, and adapt to the user's different dialogue strategies. Also, intelligent software is needed to take into account requirements for complex interaction: the users' knowledge and intentions, variation in their viewpoints and interests, and the context in which interaction takes place.

However, the more complex systems are constructed, the more complicated it is to evaluate the system 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 which can vary from quick and simple prompts to natural intuitive interaction.

The usual way of evaluating interactive systems is to interview users after they have used the system for particular tasks, collect the evaluation forms, and use a weighted function of task-based success and dialogue-based 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 (is the user satisfied). The system’s objective performance and the user’s subjective view of the usage of the system are taken into account, with the goal to maximize the objective function of user satisfaction. However, various evaluation studies show that the users seem to tolerate difficulties such as long waiting times and even mere errors, if the system is interesting and the users motivated to use it. The users thus also assess the system's qualityin 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.

Quality evaluation can be operationalised by determining quality features that represent those properties and functions of the system that contribute to the user's positive perception of the system and the service it provides. Once the features have been determined, their impact on the system performance can be quantified and compared as described above: by calculating the difference between the user’s expectations and experience before and after the use of the system, and checking how the different features have contributed to the overall change. Quality features deal with task requirements and also with the system’s capability to provide services and appropriate information that the user may find useful and reliable, in a manner that takes into account the user’s focus of attention. An important aspect in the quality evaluation of natural language interfaces is thus the system’s communicative competence: success of interaction measured in terms of dialogue and cooperation principles. Quality features can also deal with such intangible aspects as style and novelty which, however, fall outside the human-computer interaction research proper.

This Chapter has argued in favour of natural language interfaces as a means to support an intuitive approach for both designing and interacting with computer applications. Interactive systems should afford natural interaction, and thus be able to provide services that the users find intuitive to use. Based on the principles of Ideal Cooperation, it is assumed that the system's communicative competence measures the system's capability to cooperate with the user on a shared task, be it to find information, to evaluate some planning options, or just to entertain.

References

Allen, J. F., Miller, B. W., Ringger, E. K. and Sikorski, T. 1995. A Robust System for Natural Spoken Dialogue. In the *Proceedings of the 1996 Annual Meeting of the Association for Computational Linguistics (ACL'96)*, 62-70.

Allwood, J. 1976. *Linguistic Communication as Action and Cooperation.* Gothenburg Monographs in Linguistics 2. Göteborg University, Department of Linguistics.

Allwood, J., Traum, D. and Jokinen, K. 2000. Cooperation, Dialogue, and Ethics. *International Journal for Human-Computer Studies* 53(6): 871-914.

Aust, H., Oerder, M., Seide, F. and Steinbiss, V. 1995. The Philips automatic train timetable information system. *Speech Communication* 17(3): 249-262.

Austin, J. L. 1962. *How to do Things with Words.* Harvard University Press.

Bunt, H. and Girard, Y. 2005. Designing an open, multidimensional dialogue act taxonomy. In the *Proceedings of DIALOR'05*, 37-44.

Chu-Carroll, J. and Carpenter, B. 1999. Vector-based natural language call routing. *Computational Linguistics* 2(3): 361-388.

Clark, H. and Schaefer, E. 1989. Contributing to discourse. *Cognitive Science* 13:259-294

Cohen, P. R., Morgan, J. and Pollack, M. E. (Eds.). 1990. *Intentions in Communication*. Cambridge, MA: MIT Press.

Fais, L. 1994. Conversation as collaboration: some syntactic evidence. *Speech Communication* 15(3-4): 231 – 242.

Gorin A. L., Riccardi G. and Wright J. H. 1997. How May I Help You? Speech Communication, 23(1-2): 113-127.

Grice, H. P. 1957. Meaning. *The Philosophical Review* 66(3): 377-88. Reprinted in Grice (1989).

Grice, H. P. 1989. *Studies in the Way of Words*. Cambridge MA: Harvard University Press.

Grosz, B. and Sidner, C. 1986 Attention, intentions, and the structure of discourse. *Computational Linguistics* 12(3): 175-203.

Gumperz, J. and Hymes, D. (Ed.). 1972. *Directions in sociolinguistics: The ethnography of communication*. New York: Holt, Rinehart and Winston.

Hurtig, T. and Jokinen, K. 2006. Modality fusion in a route navigation system. In the *Proceedings of the IUI 2006 Workshop on Effective Multimoda Dialoguel Interfaces*, 19-24.

Kanto, K., Cheadle, M., Gambäck, B., Hansen, P., Jokinen, K., Keränen, H. and Rissanen, J. 2003. Multi-session group scenarios for speech interface design. In C. Stephanidis and J. Jacko (eds.) *Human-Computer Interaction: Theory and Practice (Part II),* volume 2, pp. 676-680, Mahwah, New Jersey, June. Lawrence Erlbaum Associates.

Jokinen, K. 1996. Cooperative Response Planning in CDM: Reasoning about Communicative Strategies. In *TWLT 11. Dialogue Management in Natural Language Systems*, eds. S. LuperFoy, A. Nijholt and G. Veldhuijzen van Zanten, 159-168. Enschede: Universiteit Twente.

Jokinen, K. 2000. Learning dialogue systems. In *From Spoken Dialogue to Full Natural Interactive Dialogue – Theory, Empirical Analysis and Evaluation*, *LREC,* ed. L. Dybkjaer, 13–17.

Jokinen, K. 2003. Natural Interaction in Spoken Dialogue Systems. In the *Proceedings of the HCI International 2003 Conference*, vol. 4, 730–734. NJ: LEA.

Jokinen, K. 2008a. *Constructive Dialogue Management – Speech Interaction and Rational Agents*. John Wiley & Sons.

Jokinen, K. 2008b. User Interaction in Mobile Navigation Applications. In *Map-based mobile services - design, interaction and usability*, eds. L. Meng, A. Zipf and S. Winter, 168-197. Springer Series on Geoinformatics.

Jokinen, K. and Hurtig. T. 2006. User Expectations and Real Experience on a Multimodal Interactive System. In the *Proceedings of Interspeech Conference*, Pittsburgh, U.S. On-line available at: http://www.ling.helsinki.fi/~kjokinen/Publ/200609InterspeechMUMSeval.pdf

Jokinen, K., Kerminen, A., Kaipainen, M., Jauhiainen, T., Wilcock, G., Turunen, M., et al. 2002. Adaptive Dialogue Systems - Interaction with Interact. In the *Proceedings of the 3rd SIGdial Workshop on Discourse and Dialogue*, eds. K. Jokinen, and S. McRoy, 64–73.

Levin, E., Pieraccini, R. and Eckert, W. 2000. A Stochastic Model of Human-machine Interaction for Learning Dialog Strategies. *IEEE Transactions on speech and audio processing* 8(1): 11-23.

McTear, M. 2004. *Spoken Dialogue Technology: toward the Conversational User Interface.* London: Springer Verlag.

Milekic, S. 2002. Towards Tangible Virtualities: Tangialities. In *Museums and the Web 2002: Selected Papers from an International Conference*, eds. D. Bearman and J. Trant. On-line available at: http://www.archimuse.com/mw2002/papers/milekic/milekic.html

Norman, D. A. 2004. *Emotional Design: Why We Love (Or Hate) Everyday Things*. Cambridge, Mass: Basic Books.

Raux, A., Langner, B., Black, A. and Eskenazi, M. 2005. Let's Go Public! Taking a Spoken Dialog System to the Real World. In the *Proceedings of Interspeech 2005*. On-line available at: http://www.cs.cmu.edu/~awb/papers/is2005/IS051938.PDF

Reeves, N. and Nass C. 1996. *The Media Equation: How people treat computers, television, and new media like real people and places*. New York: Cambridge University Press.

Rich, C. C., Sidner, C. L., Lesh, N., Garland, A., Booth, S. and Chimani, M. 2005. DiamondHelp: A Collaborative Task Guidance Framework for Complex Devices. In the *Proceedings of the 20th AAAI Conference and the 17th Innovative Applications of Artificial Intelligence Conference*, 1700-1701. Pittsburgh: AAAI Press / The MIT Press.

Rudnicky, A., Thayer, E., Constantinides, P., Tchou, C., Shern, R., Lenzo, K. et al. 1999. Creating natural dialogs in the Carnegie Mellon Communicator System. In the *Proceedings of the 6th European Conference on Speech Communication and Technology (Eurospeech-99)*, 1531–1534.

Russel, S. and Norvig, P. 2003. *Artificial Intelligence – A Modern Approach* (2nd Edition). Prentice Hall.

Sadek, D. 2005. ARTIMIS Rational Dialogue Agent Technology: An Overview. In *Multi-Agent Programming - Languages, Platforms and Applications,* vol. 15, 217-243. US: Springer.

Scheffler, K. and Young, S. 2002. Automatic learning of dialogue strategy using dialogue simulation and reinforcement learning. In the *Proceedings of Human Language Technology,* 12–18.

Searle, J. R. 1979. *Expression and Meaning: Studies in the theory of Speech Acts.* Cambridge University Press.

Seneff, S., Lau, R. and Polifroni, J. 1999. Organization, communication, and control in the GALAXY-II conversational system. In the *Proceedings of the 6th European Conference on Speech Communication and Technology (Eurospeech-99),* 1271–1274.

Wahlster, W. (ed.). 2000. *Verbmobil: Foundations of Speech-to-Speech Translation.* Heidelberg, Berlin: Springer-Verlag.

Walker, M. A., Fromer, J. C. and Narayanan, S. 1998. Learning Optimal Dialogue Strategies: A Case Study of a Spoken Dialogue Agent for Email. In the *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conf. on Computational Linguistics*, 1345--1352.

Weiser, M. 1991. The Computer for the Twenty-First Century. *Scientific American* 265(3): 94-104.

Williams, J. D. and Young, S. 2006. Partially Observable Markov Decision Processes for Spoken Dialog Systems. *Computer Speech and Language* 21(2): 393-422.

Zue, V. 1997. Conversational Interfaces: Advances and Challenges. In the *Proceedings of Eurospeech 97*, p. KN 9-18.

VoiceXML Forum. http://www.voicexml.org/

Figure captions

Figure 1 A top-level view of a dialogue system

Figure 2 General architecture of MUMS (Hurtig and Jokinen 2006)

Figure 3 A screenshot in MUMS