Exploiting Recent Research on Dialogue to Model Verbal Communication in Staging

Costanza Navarretta
Center for Sprogteknologi
costanza@cst.ku.dk

1 Introduction

Multimodal systems where the user interacts with one or more autonomous agents in a tridimensional environment must be able to interpret multimodal input from a user and react to it in a coherent way. The treatment of natural language dialogues in computer systems can be divided into three components, dialogue management, discourse interpretation and response generation. Dialogue management consists in implementing strategies that drive the dialogue, such as who can take the initiative and who controls the dialogue (the user, the system or both of them) and how the system must react to error situations. The choice of strategy depends, i.a., on the application of the system and on the degree of robustness of the available speech recognition software.

The treatment of verbal interaction in unimodal and multimodal dialogue systems is similar. In unimodal systems a dialogue management component decides how to verbally interact with a user's utterance, after having interpreted it. In a multimodal system the non-linguistic input from the user is combined and interpreted together with the linguistic input. A multimodal manager module is responsible for choosing the most adequate modalities for presenting the system's response.

Discourse interpretation and response generation are built upon models of discourse structure. Such models have been proposed in different fields, i.a. sociology, psychology, linguistics, philosophy and each of these focuses on specific aspects of discourse, relevant to the actual field. Natural language engineers and scientists from artificial intelligence have been inspired by all these models. The dialogues determining the sub-language to be handled by dialogue systems are human-to-human task-specific conversations, or dialogues collected with Wizard of OZ experiments[1]. The conversations allowed in implemented systems are very limited and follow prespecified schemata. The question which we will address is how to define strategies to implement dialogue systems which allow less bound and more "natural" dialogues between a user and agents than in existing implementations.

Although we primarily focus on discourse analysis, i.e. on the interpretation of linguistic and non-linguistic data from the user after this input has been processed by the parser and the multimodal combination components (see the chapter by Paggio and Music in this book) we also address issues of dialogue management and of response generation.

In what follows we shortly present some of the components needed by a system to interpret dialogues and generate responses. We then review a few theories for modelling these components and look at the solutions implemented in existing uni- and multimodal dialogue systems. Then we discuss how the theoretical and implemented models fit in with less schemata bound dialogues than those allowed by existing systems. Finally we propose a preliminary model to be used in an implementation of dialogue management in Staging.

In what follows we use dialogues provided by the project's dramatist group. These dialogues, from now on the *bar dialogues*, are conversations between a bartender, B, and a customer, C. The conversations are flirtatious, in that B flirts with C while he is asking what she would like to drink, whether she is feeling well and so on. These dialogues have more social implications and are less task-bound than dialogues supported by implemented systems.

2 Common Ground and Context

Most models of discourse[2] structure are based on the assumption that dialogue participants share a certain common ground and that, during conversation, they dynamically add information to this common ground. Defining and encoding the common ground, or background knowledge, of dialogue participants is quite a complex task. Background knowledge comprises linguistic and non-linguistic information, general and domain-specific (sub-language-specific) information. In a computer system the knowledge encoded is necessarily a strong simplification of the knowledge humans have and it is restricted to the sub-language and task which the system has to perform. The knowledge used by a system is both static and dynamic, the latter being the information which is changed or added to the system as a result of the interaction with the user.

One of the most important dynamic knowledge sources in a dialogue system is the so-called **dialogue history** where the context of the interaction between the user and the system (or the system's agents) is registered. The dialogue history is useful to both the analysis and the generation process. The following example (1) from the *bar dialogues* illustrates a simple case of utterance interpretation against the previous conversational context.

(1) a. B: *Generer lyset dig?*
(Does the light annoy you?)
b. C: *Det er lidt skarpt*
(It is a little bit sharp)

If we assume that B is a system agent and that C is the user, the dialogue interpreter component has to "remember" utterance (1-a) to find out that the pronoun *det* (it) in utterance (1-b) refers to the light. In the isolated example above simple restrictions on the gender of the nominals in the two utterances are sufficient to resolve the pronominal reference. But, of course, there are many cases where the situation is more complex as in the following example (2):

(2) a. *Anne gave presents to Carol.*
b. *She liked it*

If stress information is not available, sentence (2-b) is ambiguous. The preferred interpretation is that Carol liked receiving the presents. However, in a broader context where the speaker is explaining why Anne has no money, the sentence could actually mean that Anne likes giving (and buying) presents. In many cases, the interpretation of utterances requires both knowledge of the structure of the ongoing dialogue, i.a. (Hobbs 1983, Asher 1993), and knowledge of the intentions behind a speaker's utterance, i.a. (Grosz & Sidner 1986, Moore & Pollack 1992). The structure of discourse is also necessary to generate appropriate responses. In example (1), a coherent response of the bartender to the customer who utters that the light is too sharp could be replacing it with a softer light or suggesting that the customer could sit at another table. If the bartender reacted by turning sharper lights on, his response would be perceived as peculiar.

3 Discourse Structure

In this section we look at some models of discourse structure presented in the fields of natural language engineering and artificial intelligence, which have influenced the implementation of dialogue systems, the so-called Rhetorical Structure Theory, RST, (Mann & Thompson 1987), the coherence relation theory (Hobbs 1983) and what we will call the GS86 model (Grosz & Sidner 1986). The first two theories are inspired by linguistic studies investigating the relations that link clauses, sentences or other pieces of discourse to each other, for example (Halliday & Hasan 1976), while the third approach is inspired by work on dialogues in task-oriented applications and on studies of the resolution of referring expressions in these task-specific sub-languages.

3.1 Hobbs

Hobbs (1983) assumes that conversations are planned and that people interact and produce text in order to achieve certain goals. He recognises three interleaved levels of planning in conversation: the message level, the coherence level and the description level. The message level involves the main goals that motivate people to engage in a conversation, the coherence level is about how the speaker plans to present his message. The description level is concerned with the choice of lexical items, grammatical constructions and appropriate descriptions of entities and events. In modelling comprehension all the three levels must be taken into account. Hobbs' coherence theory only investigates the second level of planning, but it is part of a more general theory, according to which text interpretation means finding the most plausible explanation to a text using an abductive inference model (Hobbs, Stickel, Appelt & Martin 1993).

Texts are coherent and coherence can be characterised by means of inferable coherence relations between segments of the discourse. A coherence relation holds between sentences to a degree which depends on the salience of the axioms used to establish the relation. In Hobbs' coherence theory the basic elements of discourse are sentential units (clauses). The union of two sentential units linked by some coherence relation is also a sentential unit. The structure of discourse is built up as a binary tree, whose terminal nodes are clauses and whose non-terminal nodes are coherence relations among them.

Elaboration is one of the recognised coherence relations and is characterised as follows:

> A segment of discourse S1 is an Elaboration of segment S0 if the same

proposition P can be inferred from both S0 and S1, and one of the arguments of P is more fully specified in S1 than in S0. (Hobbs 1983, page 31)

According to Hobbs a question-answer sequence is a kind of Elaboration, as it is the case in the following example (3):

(3) a. B: *Hvad skal du have at drikke?*
(What do you want to drink?)
b. C: *En fadøl*
(a draught beer)

The information that C wants a draught beer, elaborates the question of B in the required way, i.e. it fills in the object argument for the formula

A: $drink(c, X) \land customer(c)$
so that we get
B: $drink(c, b) \land customer(c) \land draught\_beer(b)$

The restricted set of coherence relations described by Hobbs have been used in many dialogue systems. His abductive theory of inference has also inspired theories which try to combine belief and dialogue models. The main problem with the abductive inference model is that it is not very efficient and requires the collection of quite a large amount of world knowledge which the participants in a conversation are supposed to share.

3.2 Mann and Thompson

Mann and Thompson's Rhetorical Structure Theory (RST) was originally a theory for generating explanatory texts. Later on it was applied to deal with discourse interpretation. As in Hobbs' coherence theory (3.1) RST texts are divided into units, called spans, which typically consist of a clause and form the terminal nodes in an RST tree structure. Spans are related by binary rhetorical relations. The union of two spans having an RST structure is also a span. The most prominent span in each relation is called the nucleus, while the less prominent one is called the satellite. The set of nucleus-satellite relations is open. Some of the relations recognised are: Antithesis, Background, Conditional, Elaboration, Enablement, Purpose. RST only permits one relation between two spans. If a text has multiple analyses, the most plausible one in terms of the writer's perceived overall goals must be chosen. An RST analysis of sentence (4)

(4) *Jeg har også små peanuts, hvis det passer dig bedre*
(I have also small peanuts, if it is better for you)

would result into two RST spans:

1. Jeg har også små peanuts
2. hvis det passer dig bedre

Span 2 is in a **Conditional** relation with span 1, the nucleus. The graph for the obtained RST relation can be seen in figure (4). RST theory has been used in

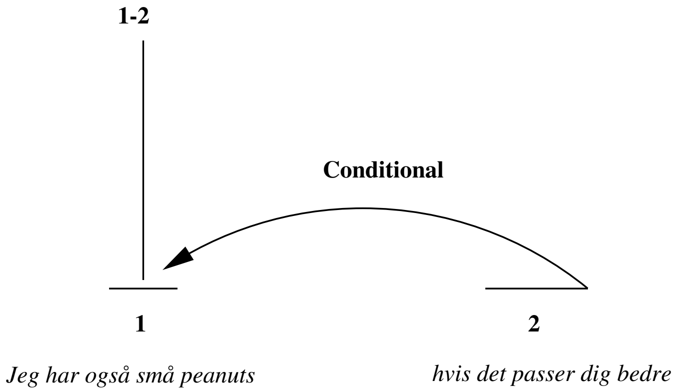

Figure 1: RST diagram for example (4)

many generation systems and in a few dialogue systems. The fact that it does not support multiple analyses of the same text has been criticised, because most texts are ambiguous (Moore & Pollack 1992, Asher & Lascarides 1994). The strong side of the theory is, in our opinion, that many of the rhetorical relations, described in the model, have been identified on the basis of the analyses of texts done by different people. We see it as a problem that cognitive aspects are not explicitly incorporated in the theory.

3.3 Grosz and Sidner

Grosz and Sidner present a model of discourse structure which combines three interacting components: the linguistic structure, the intentional structure and the attentional state. The **linguistic structure** consists of the sequence of utterances in a discourse, aggregated into so-called discourse segments, and of the relations which hold between these segments. These relations are the surface reflection of the two relations holding among elements of the intentional structure: **satisfaction-precedence** (DSP1 contributes to DSP2) and **dominance** (DSP2 dominates DSP1).

The **intentional structure** deals with the purpose of the discourse. The intention that motivates engaging in a particular discourse is the so-called discourse purpose (DP). Each discourse segment has a corresponding discourse segment purpose (DSP) which specifies how the actual segment contributes to the achievement of the DP. DP and DSPs are intentions which the speaker wants to be recognised. To determine DSPs, information from different knowledge sources must be collected such as linguistic markers (especially cue phrases), utterance-level intentions, general knowledge about actions and objects in the domain of discourse, intonation.

The attentional state coordinates the linguistic and the intentional structures and is modelled by a set of focus spaces. The attentional state dynamically records the objects, properties, relations and DSPs that are salient at each point of the discourse. The relationships among DSPs determine pushes and pops from the focus stack. The focus space constrains the range of DSPs to be considered as candidates to be related to the current DSP and the search for possible referents of nominal phrases[3].

In the GS86 model the utterances in (4) would be part of a Discourse Segment whose corresponding purpose is that of selling nuts to the customer. The main task (DP) for a bartender is, in fact, that of selling things and of serving the customer. Parallel to this overt task there is the flirtatious intention which is more difficult to formalise and to break down into subgoals.

Grosz and Sidner's theory is interesting because it relates linguistic, intentional and attentional aspects in the same framework. Grosz and Sidner's analysis of i.a. the relation between linguistic expressions and intentions, their implementation of the attentional state as a stack and the treatment of different kinds of phenomena such as anaphora resolution and interruptions, have also influenced later research on discourse and dialogue. For us the main problem with the theory is the identification of DSPs, which is not spelled out in the model. Segmenting discourse in fragments which reflect subgoals is nearly impossible when automatically treating dialogues which are not strictly task-oriented. The second problem with the theory is that the informational aspect of utterances is neglected.

3.4 Combining Intentionality and Textual Relations

There have been several attempts to combine theories of discourse structure based on rhetorical relations and theories where intentionality plays a central role in the analysis of communication. Common to these "mixed" approaches is the fact that they state that both rhetorical and coherence relations and the recognition of the intentionality behind utterances are necessary to the task of interpreting dialogues (Moore & Pollack 1992, Asher & Lascarides 1994).

Asher & Lascarides (1994), in particular, present a complex formal model based on two parallel theories. The first theory specifies how rhetorical relations influence the structure of discourse. The second theory accounts for how different pragmatic knowledge resources must be used to infer the rhetorical and coherence relations holding between discourse units. Asher & Lascarides (1998), (Asher & Lascarides 1999) extend this framework so that it integrates discourse content and cognitive states. They use non-monotonic cognitive axioms to infer rhetorical relations at the discourse level, and these rhetorical relations have a semantics which constrains the resolution of anaphora.

Asher and Lascarides' work is interesting from a theoretical point of view because they actually present a framework for combining the different approaches to discourse structure. However, their model is much more complex than those used in practical applications and has to be tested in a real-time implementation.

3.5 Concluding Remarks

What in general characterises the different theories and models which we have presented in this section is the attempt to identify the most important aspect(s) of discourse structure and the elementary units of discourse. RST concentrates on a single aspect of discourse, i.e. its rhetorical structure. This structure is useful in text generation and can influence the resolution of anaphora in text interpretation. However RST alone does not provide enough information to model dialogue in interactions where both the informational and intentional aspect are relevant.

More complex are Hobbs' and the GS86 models, in which the importance of the linguistic structure together with the intentional and cognitive structure are recognised, although only some aspects of each level have been given priority in the authors' initial proposals. Recently researchers have argued for the necessity of combining the basic aspects of these three theories.

4 Dialogue Modelling in Planning and Information Retrieval Applications

Different approaches to dialogue modelling have been followed in the field of expert systems, specially planning systems, and in information retrieval or booking systems.

4.1 Dialogue Modelling in Planning Systems

Common to all the dialogue models adopted in planning applications is that they are inspired by Searle's speech acts theory (Searle 1975). The speech acts relevant to a given application are treated and formalised as physical actions. An important aspect in the treatment of actions in expert systems is the recognition of the intentions (goals) which move an agent to perform these actions, similarly to the GS86 model. Following a tradition started with (Fikes & Nilsson 1971) actions are considered to be *operators* described by the following elements:

- a *header* containing the name of the action
- a *precondition list* storing the conditions which must apply to perform the action
- an *effect list* representing what will be true after the action
- *bodies* storing the means or subgoals/sub-actions whose achievement constitutes the performance of the action

In the last decade the interest of researchers in the field of expert systems has moved towards the cognitive aspects involved in the processing of dialogues, i.e. the formalisation of the effects of the illocutionary acts on the mental state of the conversants (Litman & Allen 1990, Cohen & Levesque 1990).

A dialogue system combining traditional linguistic modules with a plan-based dialogue model is the TRAINS system, developed at the Department of Computer Science of the University of Rochester (Allen, Schubert, Ferguson, Heeman, Hwang, Kato, Light, Martin, Miller, Poesio & Traum 1995). The TRAINS system is a natural language conversation and plan reasoning system where the user and an agent must manage in cooperation a railway transportation system.

The architecture of the TRAINS system contains natural language interpretation modules and domain task interaction modules which both interact with the dialogue manager. The dialogue model is based on the so-called conversational act theory (Traum & Hinkelman 1992) which states that speech acts must be grounded (Clark & Schaefer 1989), i.e. acknowledged, by the listener before they are added to the common ground. The dialogue manager of the TRAINS system maintains a model of its mental state and the discourse context which includes the turn, the current domain plan, grounding and plan-based focusing information. The discourse context is implemented as a stack.

The TRAINS project is interesting because it combines planning techniques with a computational linguistic approach to the treatment of dialogues. In the analysis of dialogues the central idea is that of grounding , which has been followed by many researchers since then.

4.2 Dialogue modelling in Information Retrieval Systems

Many real-time systems have followed a so-called dialogue grammar approach, i.a. the Sundial (Speech Understanding and DIALog) systems (McGlashan, Bilange, Fraser, Gilbert, Heisterkamp & Youd 1992) and the LINLIN system (Ahrenberg, Dahlbäck & Jönsson 1990). In this approach it is assumed that the structure of the chosen textual units can be used to model dialogue and that a grammar can describe the dialogue's structure. Although the intentional perspective is still considered to be important, it is not central to the processing of dialogues. The idea of using a grammar to describe the most important elements of the surface structure and the semantics of dialogue and the description of suitable parsing strategies for such a grammar has been introduced by Polanyi's (Polanyi 1988) Linguistic Discourse Model (LDM). She was inspired by the theories of discourse structure which we presented in 3 and by studies of conversations made by the American sociologists, also known as the ethnomethodologists. In LDM the basic units are sentences, here understood as clauses or false starts, hesitations and other linguistic noise. Sentences combine by rhetorical subordination and coordination relations and, on a higher, into discourse genre units, such as stories and arguments, and Discourse Adjacency units, such as question/answer pairs, compliment/response sequences.

The models behind systems based on the dialogue grammar approach differ in the way and in the granularity in which the dialogues are divided into constituent units. In SUNDIAL and in LINLIN the units are dialogue acts/moves, i.e. speech acts, enriched with dialogue relevant information. In the Sundial model more dialogue acts compose an intervention marked as initiative, reaction or evaluation. More interventions form an exchange or negotiation. More negotiations form a transaction which correspond to a DSP in the GS86 model. In LINLIN dialogue moves combined in initiative-response units form a dialogue segment. Initiative-response units are, for example, the pairs information_request/answer, information_request/assertion.

Another dimension along which these dialogue systems vary, is to what extent they integrate non-linguistic aspects, such as task descriptions and cognitive aspects. The dialogue manager of the Sundial system, for example, operates on different structures corresponding to the three levels in the GS86 model. The LINLIN system, instead, has a simpler structure.

Apart from the above mentioned diversities, the systems are quite homogeneous and treat conversations in similar types of applications and domains such as information retrieval and booking of travels and hotels. The interaction allowed in these systems is topic specific and follows a pattern which is defined on the basis of the chosen sub-language, in most cases obtained by Wizard of Oz experiments. The higher constituent units in the dialogue structure are also chosen on the basis of the current sub-language and task-definition. These units can therefore, in most cases, be used to predict the next dialogue move. The dialogue analysis is effective and the error-rate is low. Some of the linguistic generalisations in LDM are lost in the implemented systems.

4.3 Concluding Remarks

Dialogue models behind planning systems and information retrieval systems are task and sub-language specific. The former focus on the intentions behind the utterance of speech acts, the latter use the combination structure of speech acts in specific sub-languages. Interesting in these models is the study of speech acts and the inclusion into the models of phenomena studied by the ethnomethodologists such as turn-taking mechanisms and grounding phenomena.

The analysis and formalisation of the intentions and attitudes of the conversants in the plan-based theories of speech acts opens important perspectives to the analysis of dialogues. Problematic is the identification of the intentions of an agent and the efficient implementation of a plan recogniser which considers intentions. In such a recogniser the search space is extremely large and it is necessary to adopt special strategies to make it efficient in real-time systems (Jönsson 1997). This is a serious problem especially considering the fact that the described plan-based systems are quite simple and very task-oriented.

Besides the clear necessity of simplifying the discourse interpretation process in order to implement real-time systems, dialogue systems based on the dialogue grammar approach try to combine different aspects of existing discourse models. They are efficient but the interaction they allow is limited to prespecified conversation schemata.

5 Dialogue Modelling in Multimodal Systems

Most multimodal systems have the same dialogue structure as unimodal dialogue systems. Non-verbal input from the user is combined with the verbal one and a multimodal manager decides which modality/ies is/are most appropriate to express the system output, i.a. (McGlashan 1996, Johnston 1998). In multimodal systems the user can communicate by verbal (spoken or written) utterances and by pointing to objects in the virtual world with the mouse or a pen. The systems can respond with written or spoken utterances, 2D and 3D graphics. Many multimodal systems are built reusing existing dialogue systems or some of their components. Thus the dialogue models behind these systems are not new with respect to those we have discussed in 4.2. Also the applications of multimodal systems are similar to those of unimodal dialogue systems, i.e. information retrieval and ticket booking, i.a. VOYAGER (Glass, Polifroni & .Seneff 1995), WEB-GALAXY (Lau, Flammia, Pao & Zue 1997), the Waxholm system (Carlson, Hunnicutt & Gustafsson 1995), Olga (Beskow, Ellenius & McGlashan 1997).

The new aspect in multimodal systems is the possibility of combining different modalities, therefore researchers have focused on these aspects. Agents' behaviour and the way in which they use multimodality have been formalised on the basis of studies of inter-human communication. Examples are the Waxholm system where labial and eye movements are synchronised with the spoken output, the Olga system which extends this set-up to body movements, GANDALF and REA (Cassell & Vilhjálmsson 1999) where facial movements and hand gestures are integrated with verbal interaction.

Some experimental studies have been done to record the user's choice of modalities (Oviatt, DeAngeli & Kuhn 1997) and possible ways of combining modalities have been described by Martin (1997). A first framework for choosing output modality from the media and the data characteristics has been proposed in (Arens, Hovy & Vossers 1993). Because we focus on discourse interpretation, the main issue, for our future research is how multimodal communication affects discourse structure and analysis.

6 The Staging Approach

As we have seen in the previous sections there are many theoretical and applied models of discourse. The most important dimensions of variation in the models are the choice of the dialogue units and the relations which combine them, and the weight linguistic and non-linguistic aspects are given.

Model	**Unit**	**Relation**	**Structure**
RST	sentence	rhetorical relations	tree
coherence theory	sentence	coherence relations	tree
LDM	sentence questions/answers etc.	conjunction, subordination adjacency pairs	parse-tree
LINLIN	dialogue acts	initiative-response pairs	parse-tree
GS86	DS (intention) satisfaction-precedence	dominance	tree
planning systems	intention (speech act)	is-a-precondition-of /is-a-way-to	task description
TRAINS	turn-taking, grounding speech acts	argumentation	stack

Table 1: Variation of Discourse Units and Structure

The first dimension is summarised in table 1. Most models represent dialogue structure as a tree. In a few systems the dialogue structure is implemented as a stack or an array. In expert systems the interest is on the intentions behind the utterance of speech acts and the inference processes necessary to calculate goals from intentions and intentions from goals. The chosen component units are sentences, speech acts or the intentions behind a speech act. These units can be combined by textual or plan-related relations, or they can be analysed as components in adjacency pairs. Although it is generally recognised that a variety of aspects interplay in conversations, the different theories and models only focus on few of these aspects. This second dimension of variation is not as fixed as the former, because most of the theories of discourse structure have been presented as partial theories and are continuously extended to cover more aspects.

In the following sections we look at how the different theories and practices can contribute to the treatment of less schemata bound dialogues than those allowed in existing systems. Then we present the dialogue modules which will be used in Staging.

6.1 The Dialogue Structure

All dialogue models in computational linguistics and in artificial intelligence presuppose that communication is a cooperative task and we follow this assumption. Following the GS86 model we assume that linguistic, intentional and attentional levels interact in dialogue structure, but we do not base our analysis exclusively on the recognition of the intentions behind utterances, because these intentions cannot always be determined automatically, especially in not strictly task-oriented domains. Instead we choose dialogue acts, as the basic elements of the dialogue structure, but recognise the importance of intentionality and, to a certain extent, follow a task-oriented approach.

Seen from such a task-oriented point of view, dialogues imply at least two task levels. The first level, common to all dialogue types, is that of controlling and planning the dialogue[4]. The second level is the overt domain-dependent task. In the bartender domain this level includes the tasks of selling drinks/food and serving the customer which can be split up in sub-goals and plans as it is the case for tasks implemented in expert systems.

In dialogues other more covert "tasks" can be present, i.a. related to specific social relations. In the bar dialogues, for example, one such task is that of flirting with the customer. The flirt is a kind of social game whose main goal is keeping the flirt going. At this level the way and the style in which sentences are uttered and extra-linguistic factors are more important than what is said.

To treat the dialogue-task level we follow the dialogue grammar approach defining initiative-response pairs and sub-language specific pairs. Few coherence and rhetorical relations can also be used. The domain-specific task can be treated in a task module, where a task is divided into subtasks which are connected to specific linguistic expressions (here dialogue acts). The components at this level correspond to the discourse segments in the GS86 model. They are used in the interpretation phase to keep track of entities relevant at the current point of the interaction for i.a. resolving referring expression and treating interruptions. The task structure is also used when the system has to plan the next move of the agent.

More social related tasks can require the generation of specific types of sentences (in a flirt situation an agent's utterances could be ambiguous). Use of textual relations as in RST and in the coherence theory and commonsense knowledge could be useful at this level.

6.2 Dialogue Modules

A **dialogue manager** module is in charge of interpreting the user's multimodal response against the context of the on-going interaction and of planning verbal reactions. The interpretation is made in a **dialogue analysis component** which receives a semantic representation of the user's input from the multimodal combination component (see Paggio's and Music's paper in this volume). The dialogue analysis component is in charge of resolving referring expressions and of interpreting the output in the current interaction. The interpreted message is then passed to the multimodal manager which is responsible for choosing an appropriate response from the agent. If the chosen response includes verbal output, the dialogue manager passes an appropriate semantic representation of the utterance to be output to a **generation component** and then adds this representation to the dialogue history. A component describing the task structure is consulted by the dialogue manager during both interpretation and generation. In what follows, we focus on the interpretation process. In unimodal dialogue systems the universe of discourse, i.e. the entities the user and the system can verbally refer to, consist exclusively of the entities that the user and the system/agents are expected to know at the beginning of the interaction (common background defined by the application sub-language and task) and the entities that are explicitly introduced during the interaction.

In multimodal communication, where the user interacts with agents in a 3D virtual world, the objects and agent/s in this world are also part of the common ground. The on-going changes in the virtual world must also be recorded and the user and the agent/s can refer to the entities of the 3D-world verbally and with pointing acts.

The dialogue manager keeps track of the dialogue in the **dialogue history**, a representation of the user's and the agent's conversational contributions. The dialogue manager must also have access to an other dynamic data structure recording the entities (objects, agents) currently present in the 3D world. The dialogue analysis component receives a simple semantic representation of the combined user's verbal and non-verbal input, together with information about the sentence-type, such as whether the sentence is a question, an assertion, an interjection[5].

The dialogue history is composed of dialogue units. A dialogue unit is a dialogue act, i.e. a speech act to which is added information about the dialogue "situation", using Barwise and Perry's terminology (Barwise & Perry 1983). In most cases a sentence corresponds to a speech act, but in some cases (e.g. complex sentences) it must be split up into more units. The dialogue analysis component adds to the original semantic formula the information about the utterance situation. If the representation corresponds to a single speech act it will be added to the dialogue history, represented as a tree, in conformity with LDM. If the sentence representation corresponds to more speech acts, it is split up before the resulting acts are added to the dialogue history. A dialogue act is part of an initiative-response pair and more pairs will combine into a transaction (a discourse segment in the GS86 model). Transactions in the bartender domain are, for example, selling a drink, selling food, asking how the customer is feeling.

The insertion and deletion of dialogue acts into/from the dialogue tree is made according to these structuring elements. For example a transaction will start a new subtree, whose daughters are initiative-response pairs. When a transaction is terminated, the corresponding subtree can be deleted.

We formalise dialogue using a variation of Discourse Representation Theory, **DRT** (Kamp 1981, Kamp & Reyle 1993). **DRT** is the first of a number of semantic theories based on a dynamic notion of meaning, i.e. where the meaning of a sentence is considered a function from a context to a context. DRT describes how to construct semantic representations, called *Discourse Representation Structures*, DRSs, from natural language (English) sentences.

As an example we give the graphical DRS for utterance (1-a) (*Does the light annoy you*).

(K0)

e1,x,y
light(x) *customer(y)* *annoy(e1,x,y)*

The DRS is added the information that the sentence has been uttered by the agent and that it is a yes/no-question.

(K1)

e1,x,y
light(x) *customer(y)* *annoy(e1,x,y)*

utterer: agent1; sentence_type:Yes-No-question

A "yes-no-question" in the present example is an "information-request" act that in most cases is followed by an "answer" act (4.2). Here we assume that K1 starts a new serving transition about the "light". A transition node is added to the existing dialogue tree and a daughter-node, marked as info-request/answer, is then attached to it. DRS K1 is added to this node.

The customer's response in(1-b) (*It is a little bit sharp*) can be represented by the enriched DRS K2.

(K2)	*s1,x,z*
	sharp(s1,z) *little(s1)* *z=x*

utterer: customer; sentence_type:assertion

The dialogue analysis component resolves the reference of pronoun *it* to the *light* i.a. on the basis of gender information from the parser. The problem, at this point, is to "understand" how K2 is answering K1, i.e. what is the "occasional" meaning of the utterance, using a Gricean expression. The dialogue model is based on the assumption that communication is a cooperative enterprise and that the user's intention in uttering (1-b) is to provide a coherent response to the agent's question. To interpret K2 as a coherent answer to K1, the dialogue manager needs more knowledge. This knowledge, called lexical semantic knowledge, could, in this particular case, be the fact that the adjective *sharp* modifying the nominal *light* means "strong" and that *sharp light* can be annoying. At this point the system can infer that the customer, by answering that the light is a little bit sharp, really means that she is annoyed by the light. This is in line with Hobbs' coherence theory and his abductive interpretation theory, but it involves also the notion of intentionality as it is presented in the GS86 model and in models behind planning systems. It most be noted that encoding all the knowledge necessary to make this kind of inferences is, by no means, a trivial task.

In the two extended DRSs K1 and K2 we have not included information about time and space. This information, comprising reference to the time and place in which an utterance has been performed, together with the time and place to which the current dialogue act refers, must be added to the representation of each utterance. Time and space information in the actual dialogue unit come from the parser (tense information and temporal and spatial adverbials). Spatial information must be related to the current 3D world. This information is used to interpret deictic expressions. If we add this information to DRS K1 we get K3. In K3, **t** stands for time, **p** for place, **n** stands for "now" and **b** stands for "in-the-bar"[6].

(K3)

e1,x,y,t,p,n,b
t=n *p=b* *light(x)* *customer(y)* *annoy(e1,x,y,t,p)*

utterer: agent1; sentence_type:Yes-No-question, place=p, time=n

The time and place to which the dialogue act refers are, of course, not always the same as the time and place in which the act is uttered.

If no deictic gestures combine with the utterance *Does the light annoy you*, and if no reference to a *light* could be found in the preceding context, the definite nominal phrase *the light* must be interpreted using background knowledge of the 3D world, i.e. descriptions related to the light entity/ies in the virtual world. If the agent points to a specific light in the bar, while producing the utterance, the information of the gesture's type and direction, combined with the information of the entities in the 3D world, is used to interpret the nominal phrase as referring to the light which the agent is pointing to. Algorithms for resolving referring expression are usually restricted to anaphora and, in some cases, to temporal deictic expressions. They must be extended to treat other deictics.

7 Conclusion

To analyse and coherently respond to user multimodal input, where language is one of the modalities involved, is a very complex task. In this paper we have discussed how discourse structure can be modelled in a multimodal scenario. In particular we have looked at how existing discourse structure theories and models can contribute to the formalisation of complex dialogues between a user and an agent in 3D virtual worlds and we have proposed an approach to be used in the Staging project. We have also presented modules to process dialogues in a multimodal set-up. In this set-up the representation of the context of discourse must be extended to comprise the entities (agents and objects) in the virtual world and the changes in their status during the interaction. Although we have mainly focused on the analysis process we have also discussed issues relevant to dialogue management and answer generation.

The theories and models of discourse structure which we have discussed are partial, focusing on few aspects of discourse. We argue that some of the aspects and strategies from the theories must be combined for tackling more natural (human-like) dialogues than those allowed in many implemented systems. In particular we propose a preliminary mixed approach, where we combine the three level analysis presented by Grosz and Sidner with the analysis of speech acts and the use of a dialogue grammar typical of implemented models. We also argue for the importance of using commonsense background knowledge and of adopting a subset of textual relations (RST and coherence theory) in the interpretation and generation process.

Although we do not choose the intentions behind utterances as basic discourse units, because it is not clear how to extract them automatically, we follow a task-oriented approach and recognise different task levels in dialogues: a dialogue-specific, a domain-specific and a social-related task level. The presented model foresees actions at each level.

8 Footnotes

[1]In these experiments the users think that they are interacting with a real computer system, but, in reality, the system, or some of its components, is/are simulated by humans (Newell, Arnott, Carter & Cruickshank 1990).

[2]In the literature "discourse" refers to both written and spoken texts/dialogues, including multimodal input, such as gestures and facial movements. We use the term with this broad reading. Also text must be understood with its wide meaning of "written or spoken passage" (Sinclair, Hanks, Fox, Moon & Stock 1987)[p. 1510].

[3]A focus stack is used to resolve referents between discourse segments, which exhibit *global coherence*. The so-called centering theory has been proposed to resolve references to nominals among utterances in a discourse segment which are assumed to have *local coherence*, i.a. (Grosz, Joshi & Weinstein 1986, Kameyama 1998).

[4]This dialogue task level is called domain-independent plan in (Litman & Allen 1990) and dialogue control act level in (Bunt 1995).

[5]Sentence must be understood here as words and/or gesture that can be interpreted as complete propositions.

[6]A discussion of which theory should be chosen for representing temporal and spatial aspects is not in the scope of the present paper, but see i.a. (Barwise & Perry 1983, Kamp & Reyle 1993, Asher 1993).

References

Ahrenberg, L., Dahlbäck, N. & Jönsson, A. (1990), Discourse Representation and Discourse Management for a Natural Language Dialogue System, *in* 'Proceedings of the Second Nordic Conference on Text Comprehension in Man and Machine', Täby, Stockholm.

Allen, J. F., Schubert, L. K., Ferguson, G., Heeman, P., Hwang, C., Kato, T., Light, M., Martin, N., Miller, B., Poesio, M. & Traum, D. R. (1995), The TRAINS project: a case study in building a conversational planning agent, *in* 'Journal of Experimental and Theoretical Artificial Intelligence', Vol. 7, pp. 7–48.

Arens, Y., Hovy, E. & Vossers, M. (1993), On the Knowledge Underlying Multimedia Presentations, *in* M. T. Maybury, ed., 'Intelligent Multimedia Interfaces', AAAI Press, Cambridge, pp. 280–306.

Asher, N. (1993), *Reference to Abstract Objects in Discourse*, Vol. 50 of *Studies in Linguistics and Philosophy*, Kluwer Academic Publishers, Dordrecht, the Netherlands.

Asher, N. & Lascarides, A. (1994), Intentions and Information in Discourse, *in* 'Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics', Las Cruces, USA, pp. 34–41.

Asher, N. & Lascarides, A. (1998), 'Questions in Dialogues', *Linguistics and Philosophy* **23**(3), 237–309.

Asher, N. & Lascarides, A. (1999), Cognitive States, Discourse Structure and the Content of Dialogue, *in* 'Preproceedings of Amstelogue'99. Workshop on the Semantics and Pragmatics of Dialogue', Amsterdam University.

Barwise, J. & Perry, J. (1983), *Situations and Attitudes*, MIT Press, Cambridge, MA.

Beskow, J., Ellenius, K. & McGlashan, S. (1997), Olga - A Dialogue System with an Animated Talking Agent, *in* 'Proceedings of EUROSPEECH', Rhodes, Greece.

Bunt, H. (1995), Dialogue control functions and interaction design, *in* R. J. Benn, M. Baker & M. Reiner, eds, 'Dialogue in Instruction', Springer Verlag, Heidelberg, pp. 197–214.

Carlson, R., Hunnicutt, S. & Gustafsson, J. (1995), Dialog management in the Waxholm system, *in* 'Proceedings of Spoken Dialogue Systems', Vigsø.

Cassell, J. & Vilhjálmsson, H. (1999), Fully embodied conversational avatars: Making communicative behaviors autonomous, *in* 'Autonomous Agents and Multi-Agent Systems', Vol. 2,1, pp. 45–64.

Clark, H. H. & Schaefer, E. F. (1989), 'Contributing to discourse', *Cognitive Science* **13**, 259–294.

Cohen, P. R. & Levesque, H. J. (1990), Rational Interaction as the Basis for Communication, *in* P. R. Cohen, J. Morgan & M. E. Pollack, eds, 'Intentions in Communication', System Development Benchmark Series, The MIT Press, pp. 221–255.

Fikes, R. & Nilsson, N. (1971), 'STRIPS: A New Approach to the Application of Theorem Proving to Problem Solving', *Artificial Intelligence* **2**, 189–208.

Glass, J., Polifroni, J. & .Seneff, S. (1995), 'Multilingual Spoken-language Understanding in the MIT Voyager System', *Speech Communication* **17**, 1–18.

Grosz, B. J. & Sidner, C. L. (1986), 'Attention, Intentions, and the Structure of Discourse', *Computational Linguistics* **12**(3), 175–284.

Grosz, B., Joshi, A. K. & Weinstein, S. (1986), Towars a computational theory of discourse interpretation, Unpublished ms.

Halliday, M. & Hasan, R. (1976), *Cohesion in English*, Longman, London.

Hobbs, J. R. (1983), Why Is Discourse Coherent?, *in* F. Neubauer, ed., 'Coherence In Natural-Language Texts', Vol. 38 of *Papers in Textlinguistics*, Helmut Buske Verlag Hamburg, pp. 29–70.

Hobbs, J., Stickel, M., Appelt, D. & Martin, P. (1993), 'Interpretation as Abduction', *Artificial Intelligence* **63**(1-2), 69–142.

Johnston, M. (1998), Multimodal Unification-based Grammars, *in* 'AAAI Workshop on Representations for Multimodal Human-Computer Interaction'.

Jönsson, A. (1997), 'A Model for Habitable and Efficient Dialogue Management', *Natural Language Engineering* **3**(2 & 3), 103–122.

Kameyama, M. (1998), Intrasentential centering: A case study., *in* M. Walker, A. Joshi & E. Prince, eds, 'Centering Theory in Discourse', Oxford University Press, Oxford, U.K., pp. 89–112.

Kamp, H. (1981), A Theory of Truth and Semantic Representation, *in* J. A. G. Groenendijk, T. M. V. Janssen & M. B. J. Stokhof, eds, 'Formal Methods in the Study of Language', Vol. Part 1, Foris, pp. 277–322.

Kamp, H. & Reyle, U. (1993), *From Discourse to Logic: Introduction til Modeltheoretic Semantics of Natural Language, Logic and Mind*, Kluwer Academic Publishers.

Lau, R., Flammia, G., Pao, C. & Zue, V. (1997), WebGALAXY: Beyond Point and Click - A Conversational Interface to a Browser, *in* M. Gemeresereth & A. Patterson, eds, 'Proceedings of the Sixth International World Wide Web Conference', Santa Clara, pp. 119–128.

Litman, D. J. & Allen, J. F. (1990), Discourse Processing and Commonsense Plans, *in* P. R. Cohen, J. Morgan & M. E. Pollack, eds, 'Intentions in Communication', System Development Benchmark Series, The MIT Press, pp. 365–388.

Mann, W. & Thompson, S. A. (1987), Rhetorical Structure Theory: Description and Construction of Text Structures, *in* G. Kempen, ed., 'Natural Language Generation', number 135 *in* 'NATO ASI', Martinus Nijhoff Publishers, pp. 85–95.

Martin, J.-C. (1997), Towards "Intelligent" Cooperation between Modalities. The Example of a System Enabling Multimodal Interaction with a Map, *in* 'Proceedings of the IJCAI'97 Workshop on "Intelligent Multimodal Systems"', Nagoya, Japan.

McGlashan, S. (1996), Towards Multimodal Dialogue Management, *in* 'Proceedings of 11thTwente Workshp on Language in Texhnology', Enschede, The Netherlands.

McGlashan, S., Bilange, E., Fraser, N., Gilbert, N., Heisterkamp, P. & Youd, N. (1992), Dialogue Management for Telephone Information Systems, *in* 'Conference on Applied Language Processing', Trento, Italy.

Moore, J. D. & Pollack, M. E. (1992), 'A Problem for RST: The Need for Multi-Level Discourse Analysis', *Computational Linguistics* **18**(4), 537–544.

Newell, A. F., Arnott, J. L., Carter, K. & Cruickshank, G. (1990), 'Listening Typewriter Simulation Studies', *International Journal of Man-Machine Studies* **33**, 1–19.

Oviatt, S., DeAngeli, A. & Kuhn, K. (1997), Integration and Synchronization of Input Modes during Multimodal Human-Computer Interaction, *in* A. Press, ed., 'Proceedings of Conference on Human Factors in Computing Systems', New York.

Polanyi, L. (1988), 'A Formal Model of the Structure of Discourse', *Journal of Pragmatics* **12**(5/6), 601–638.

Searle, J. R. (1975), A Taxonomy of Illocutionary Acts, *in* K. Gunderson, ed., 'Language, Mind, and Knowledge', Vol. VII of *Minnesota Studies in the Philosophy of Science*, University of Minnesota Press, Minneapolis, pp. 344–369.

Sinclair, J., Hanks, P., Fox, G., Moon, R. & Stock, P., eds (1987), *Collins Cobuild English Language Dictionary*, HarperCollins Publishers, London.

Traum, D. R. & Hinkelman, E. A. (1992), 'Conversation Acts in Task-oriented Spoken Dialogue', *Computational Intelligence* **8**(3), 575–599.