

# Multimodal Communication in the Virtual Farm of the Staging Project

**Patrizia Paggio**

Center for sprogteknologi  
Njalsgade 80  
DK-2300 Copenhagen S  
+45 3532 9072  
patrizia@cst.dk

**Bart Jongejan**

Center for sprogteknologi  
Njalsgade 80  
DK-2300 Copenhagen S  
+45 3532 9072  
bart@cst.dk

## 1 Abstract

This paper describes the multimodal communication components of a 3D virtual world application that is being developed within the Danish research project Staging. The application is a farm scenario in which the user can interact with an autonomous farmer agent via an avatar to participate in simple conversations relating to life and work on the farm. Communication is multimodal in that the agent can respond to verbal input combined with a limited number of hand gestures, and it is embedded in a mixed-initiative dialogue model based on dialogue obligations and goals. In addition to interacting with the user's avatar, the agent reacts to other stimuli in the world, and its course of action – for instance whether or not to cooperate with the user – will ultimately depend on the relevance of specific behaviours at any point in time.

### 1.1 Keywords

Multimodal interaction, 3D virtual world applications, mixed-initiative dialogue.

## 2 Introduction

This paper describes work in progress at the Center for Sprogteknologi (CST) to develop the multimodal communication components of a 3D virtual world application. This work is part of an interdisciplinary Danish research project – the Staging project – the purpose of which is to investigate the nature and application potential of 3D virtual worlds, and their impact in areas such as theatre and entertainment (see Qvortrup (2000) [7] for an overview). Currently, the project is working on a farm scenario populated by a farmer and a number of animals. Interaction between the farmer and the animals builds on an interactive improvisation defined for the EU-funded Puppet project in Klasen *et al* (2000) [3]. The farmer's main goal is to establish order on the farm by herding and feeding the animals, whilst one of them, the black sheep, continually tries to disturb this order and interfere with the farmer's plans. In the present study we focus on the herding and feeding scenes foreseen by the improvisation rather than attempting to address its more dramatic aspects. Within this

limited scenario, the user can interact with the world via an avatar acting as the farmer's help.

Two characteristics of the application are challenging from the point of view of how communication between the user and the agents can be modelled. First of all, we want communication to be multimodal: the user must be able to talk, use gestures and write if needed. Secondly, interaction must support the fact that the virtual world is not static, and user and agents must be able to relate to such changes while communicating with one another. This means that communication cannot be strictly planned, and a balance must be found between the agents' wish to engage in a dialogue and their autonomous behaviour.

Our approach to multimodality is based on a feature-based multimodal parser, which pairs speech and gesture inputs during chart initialisation. The grammar merges then the multimodal information into a unified semantic representation where possible. The semantic representation is then interpreted by a communication manager according to a model aimed at dealing with mixed-initiative dialogue. The two building blocks of the model are dialogue obligations modelling valid speech act sequences, and dialogue goals modelling the agents' domain-oriented initiatives in the dialogue.

## 3 The Staging Virtual World

At the core of the Staging virtual world is the Virtual Environment (VE) server in charge of simulating the world, and to which relate an arbitrary number of autonomous agents as well as the user's avatar (for more details, see Paggio *et al* (2000) [6]). The VE provides the agents with sensory information that enable them to move around in the world and react to the presence of other agents, and it processes requests from agents, for example a request to move an object, produce a sound or play an animation.

To choose among competing behaviours (only some of which may be the result of a verbal exchange with the user's avatar), agents are equipped with a set of Releasing Mechanisms (as described in Madsen and Granum (2000)

[4]) that compute the relevance of specific behaviours at any point in time. Agents continuously attempt to carry out the most relevant behaviour, i.e. the behaviour that displays the strongest activation level. If for example the user (via an avatar) asks the farmer to feed a certain animal, say cow\$3, Releasing Mechanisms within the farmer will increase the relevance of the *feed object cow\$3* behaviour. Whether the agent chooses to perform this action depends on the level of relevance of other behaviours at that moment in time – the farmer may be hungry, tired, or disobedient and thus other behaviours may be more relevant than *feed object cow\$3*. In fact, engaging in a conversation with the user is a behaviour in itself: to win over others, it must have a high activation level. The multimodal communication components are being developed for the Staging VE. Currently, however, they are tested against a VE mock-up with reduced functionality.

#### 4 Multimodal communication

In the architecture we have defined, each agent (but not the user's avatar) will be equipped with its own multimodal analysis component. The agents that are close enough to hear the avatar speak or to see its movements or both can react to the input in different ways. For instance a dog agent, depending on how anthropomorphic it is meant to be, will either be able to react to very few linguistic utterances (*Move!*, *Sit!*, etc.), or to participate in a conversation. A humanoid agent, for example the farmer, has more complex communicative skills. Its parser can analyse speech input and gesture information. Input is provided via a microphone, a touch screen and a data glove. The speech input is sent to an off-the-shelf speech recogniser (Dragon NaturallySpeaking Professional) which provides word-based, continuous speech analysis trainable to different speakers' voices and accents. The gesture input is sent to a gesture recogniser developed by Karin Husballe Munk at CVMT (Aalborg University).

In the interface we are developing, speech can be combined with (a limited number of) deictic, iconic and turn-taking gestures following i.a. Cassell and Prevost (1996) [2]. Interestingly, these three gesture types interact with speech in different ways, as shown by the following examples:

- a) Please, feed the animals/them.  
(while pointing at a group of pigs)
- b) Move the smaller cow  
(with a gesture meaning *small*)
- c) No, I meant the other animal.  
(while waving with a hand)

In example (a), the pointing gesture is associated either with a full nominal phrase or a pronominal placeholder. It adds

referential information that must be compatible with the semantics of the rest of the speech act. In (b), the gesture adds a different kind of semantic information, which can usefully be represented as a feature contributing to the semantics of the noun *cow*. Object references coming from pointing gestures, or size information coming from iconic gestures are inserted into the chart and, if syntactically and semantically compatible, merged with the speech input. Contradictions and ambiguities are resolved where possible (see Paggio and Jongejan (2000) [5]). Both pointing and iconic gestures are used by the system to resolve nominal phrase reference, and take precedence over reference resolution based on the dialogue history.

Example (c), however, is different, in that it does not directly contribute meaning to the utterance, and is rather to be interpreted as an independent speech act. The turn-taking action should be sent directly to the VE: if the agent sees the gesture, it should stop to listen to the user's next utterance, maybe after having interrupted an ongoing action. At the moment, turn-taking gestures are recognised but not used in the VE. Turn-giving gestures, on the other hand, cause the agent to take a dialogue initiative.

#### 5 The Communication Manager

The humanoid agent is equipped with a communication manager feeding interpretations of the user's communicative acts to the agent. The communication manager is in charge of managing the dialogue with the user in accordance with a dialogue model. This implies interpreting the user messages with respect to an ongoing dialogue, deciding on the agent's dialogue moves, and building a dialogue history. Our model aims at accommodating the flexibility required by the application by combining dialogue goals arising from the scenario with dialogue obligations expressed in terms of sequences of speech acts. Dialogue goals and dialogue obligations are modelled based on well-known approaches. However, we believe these have been combined in an interesting way to be applied in a world of autonomous agents in which communicating is one among a set of competing behaviours.

##### 5.1 Dialogue goals

Examples of informally specified dialogue goals from the farm scenario – some very general, some more specific – are *negotiate whether some animal should be moved*, *negotiate whether some animal should be fed*, *talk about the weather*, and *ask for object reference*. Dialogue goals are action plans referring to a domain-relevant set of action templates in the spirit of Badler *et al* (1999) [1]. Such templates specify for a given action which semantic objects are involved, the corresponding attribute name in the semantic representation output by the parser, relevant preconditions if any precondition holds, etc. Templates for

*feeding* and *moving*, for example, may look as follows (somewhat simplified):

```
FeedAction(Topic=Feed,
           Animal=<arg3>,
           Food=<arg2>,
           Tool=<instr>)
```

```
MoveAction(Topic=Move,
           Object=<arg2>,
           Modus=<place>,
           Pace=(quickly|slowly|normal))
```

Dialogue goals are generated dynamically either to reflect changes in the VE as experienced by the agent, or to help the agent react properly to a request by the user. For example, if the farmer becomes aware of the fact that a pig is hungry, maybe because a certain amount of time has elapsed since he last fed it, his communication manager will create a dialogue goal urging him to ask the user whether the pig in question should be fed. Or if the user (U) asks the agent (A) to feed the cows, the agent will ask which food they should be fed with as shown in the following transcript (where *cow\$2* is the system internal

representation of the reference provided by a pointing gesture):

```
U: Hi feed an animal please.
A: Which animal shall I take?
U: Feed cow$2|that cow.
A: Which food shall I take?
U: An apple.
```

Of course, *dialogue* goals are not the only goals an agent will have. The exchange above, for example, will yield the goal of actually feeding the cow. Whether the action is then carried out will depend on facts external to the agent's will, such as in this case the cow's hunger level.

## 5.2 Dialogue obligations and dialogue trees

Dialogue obligations are stated as a set of condition/obligation pairs following the model described in Traum and Heeman (1996) [8]. For instance, a question is followed by an answer, an assertion by a confirmation or a contradiction, and so on. Dialogue obligations are used by the communication manager to produce a correct reaction to a speech act by the user as well as to interpret a dialogue move by the user either as closing an open dialogue segment, or as initiating a sub-dialogue.

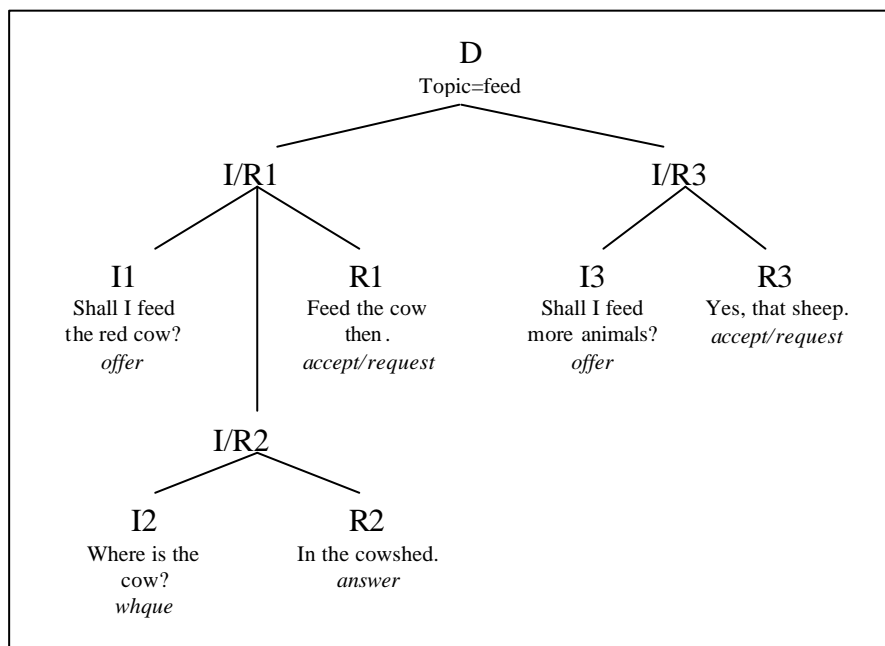


Figure 1: An agent-initiated dialogue

An example of the analysis this mechanism yields for a

short construed dialogue fragment is shown in Figure 1<sup>1</sup>.

<sup>1</sup> Pronominal reference is not dealt with yet, therefore pronouns are avoided even though they would make the dialogue more natural.

In addition to a string corresponding to the message, each terminal node in the dialogue tree contains a semantic and pragmatic interpretation of it including speech act type. In the figure, the speech act is shown in italics, together with a node index and the message. Following the DAMSL annotation scheme described in Core and Allen (97) [9], we make a distinction between *backward-looking* and *forward-looking* speech acts. A backward-looking speech act constitutes a reaction to a preceding utterance, while a forward-looking one elicits a response (either a verbal response or an action). A given dialogue move may well have both a backward-looking and a forward-looking function. From the communication manager's perspective, forward-looking speech acts specify rules conditions, while backward-looking speech acts are expected reactions to these conditions – our dialogue obligations. If a different speech act occurs instead of the expected obligation, a sub-dialogue is opened.

In the dialogue example under consideration, the first initiative (I1) is taken by the agent. It is a forward-looking speech act of type *offer*, which the communication manager interprets as opening the first initiative-response pair (I/R1) of dialogue D (see Jönsson (1997) in [10]). The topic of the dialogue is defined as *feed*. Instead of answering the agent's question directly by either accepting or rejecting the offer, the user asks for more information. The user's move is therefore interpreted as a forward-looking *whque(stion)* opening a new initiative (I2), to which the agent replies conveniently with an *answer* move (R2). This closes the sub-dialogue consisting of the second I/R pair. The next user's move is interpreted as a reply to the open I/R1, since it can be interpreted as a (backward-looking) *accept(ance)* of the agent's original offer. Note that this move also has a forward-looking function of type *request*. We shall return to this point below. Before satisfying the user's request of feeding the red cow, the agent asks whether there are other possible object candidates for the feeding action. This initiative, the *offer* in I3, is generated by a dialogue goal based on the relevant action template. The subsequent response by the user closes I/R3 as well as the whole dialogue.

### 5.3 Implied and coerced speech acts

Even in simplified and domain-oriented dialogues of the type we are dealing with, we know that the same message can be assigned several speech act labels, and that responses do not always seem to obey our limited set of condition/obligation rules. The first issue will be left undiscussed here. As for the second, we would like to examine two cases in which the rules can be relaxed in a systematic way to make the dialogue model more flexible.

One way to relax the condition-obligation rules is by allowing responses to be implied. For example, let us consider the following dialogue:

```
A: Hi.  
U: Feed the animals.
```

Since the user does not answer the agent's greeting right away, the agent should not expect to be greeted back later. In other words, after the first utterance of the human participant U, the agent must consider the greeting procedure complete. If we stick to our model of initiative-response pairs, then we have to allow for implied responses. In the example, the *Feed the animals* request initiates a new initiative-response pair. If we do not want to regard this I/R pair as part of a sub-dialogue of the greeting séance (which would seem unnatural), then we have to conclude the greeting I/R pair with an implied *greeting* response.

Another case in which the condition/obligation rules cannot be applied in a straightforward way is the following apparently unproblematic exchange.

```
U: Feed the animal.  
A: Which animal shall I take?  
U: Feed that cow$1|cow.  
A: Which food shall I take?
```

The second request (*feed that cow\$1/cow*) is not a new feeding request and must not be interpreted as an initiative opening a new sub-dialogue. On the contrary, it must be accepted as a valid response to the question (*which animal shall I take?*), and an elaboration of the action under discussion. In the current implementation a request can be coerced into an answer to a question the agent is asking to fill in the semantic arguments of an action under negotiation (as in this example), or an acceptance of such an action (as in the example in Figure 1). This, however, only happens if there is an action on the action stack that can be enriched by the information provided by the request. More precisely, for the request to be assigned the expected backward-looking speech act type, there must be semantic equality between the verb of the request and that of the pending action (*feed*), as well as semantic compatibility between their arguments (a cow is an animal).

## 6 The prototype

The prototype consists of several programs that are chained together using CORBA and Microsoft's COM as communication protocols (see Figure 2).

Dragon Naturally Speaking runs under the Windows 98 operating system, as does the Speech Recognition (SR) wrapper. The Staging VE runs under the IRIX (Silicon Graphics) operating system. The simplified VE developed at CST runs under Windows 95/98. The Communication Manager and the Multimodal Parser are platform independent and have been tested under Sun Solaris and Windows 95/98. Thanks to the CORBA interfaces all these programs can be distributed over several computers.

The SR-wrapper, the VE mock-up, the Communication Manager and the Multimodal Parser are all written in C++. A considerable part of the functionality of some of these modules is defined in external text files: the parser loads lexicon, grammar and semantic mapping rules at start-up, whereas the Communication Manager loads and reloads script files describing virtually all of its functionality. The ability to reload the script files while the prototype is running allows for a rapid development of this component.

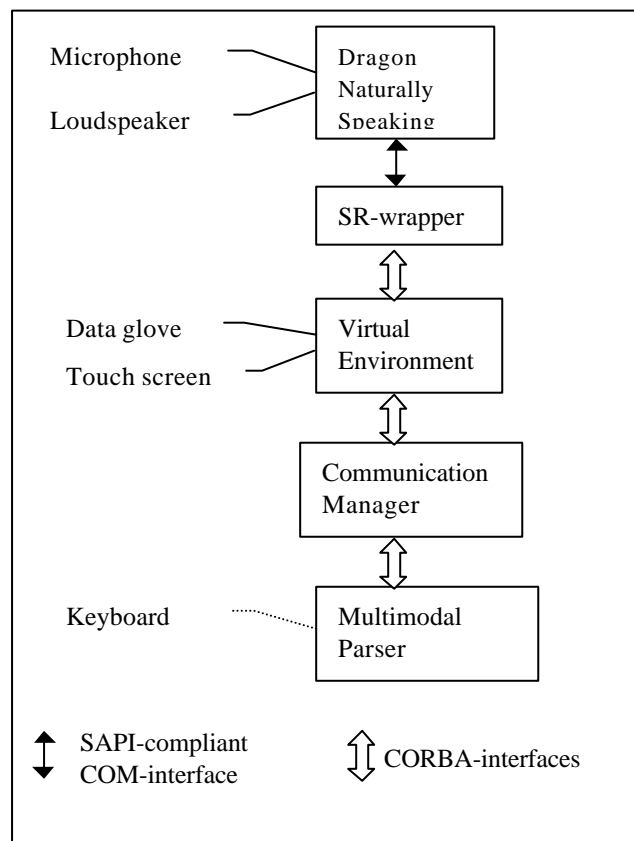


Figure 2: The prototype

## 7 Conclusion

To meet the requirements of a natural and non-task-oriented communication in a virtual world, we have developed an experimental platform where the user can use speech and gestures to interact with a humanoid

agent, and user and agent can engage in mixed-initiative dialogues dealing with work on a virtual farm.

As long as it is engaged in the dialogue, the agent will always try to satisfy a dialogue obligation in response to a user initiative. However, it can also pursue a different dialogue goal, or even disregard the user altogether to perform an action in the world. We have seen that the choice is meant to depend on which of the behaviours has the strongest activation level. Relevant factors may be the agent's personality. i.e. how polite it is (obliging a user initiative will then have high priority); a prolonged silence on the part of the user; the agent's sensitivity towards external stimuli such as the animals' restlessness (which shows their hunger), as well as the strength of these stimuli. In the VE mock-up that is part of the CST prototype, however, the potential variation offered by this set-up is not fully exploited. Thus, it is our plan to investigate how dialogue and other behaviours can be made to complement each other in a more interesting and entertaining way. Future efforts will also be devoted to further developing the domain-oriented dialogue goals, and to studying the interaction between discourse and deictic reference.

## REFERENCES

1. Badler, N.I., Bindiganavale, R., Allbeck, J., Schuler, W., Zhao, L. and Palmer, M. (2000) Parameterized action representation for virtual human agents, in J.Cassell, J.Sullivan, S.Prevoost and E.Churchill (eds) *Embodied Conversational Agents*, MIT Press, 256–284.
2. Cassell, J. and Prevoost, S. (1996) Distribution of semantic features across speech & gesture by humans and machines, in *Proceedings of the Integration of Gesture in Language and Speech*.
3. Klasen, M., Szatkowski, J. and Lehmann, N. (2000) The black sheep – interactive improvisation in a 3D virtual world, in *Proceedings of the i3 Annual Conference*, Jönköping, 13–15.
4. Madsen, C.B. and Granum, E. (2001) Aspects of interactive autonomy and perception, in L.Qvortrup (ed.) *Inhabited 3D Virtual Spaces*, Springer Verlag, chapter 3, 182–208.
5. Paggio, P. and Jongejan, B. (2000) Representing multimodal input in a unification-based system: the Staging project, in *Proceedings of the Workshop on Integrating Information from Different Channels in Multi-Media Contexts at ESSLLI 2000*, 24–31.
6. Paggio, P., Jongejan, B. and Madsen, C.B. (2000) Unification-based multimodal analysis in a 3D virtual world: the staging project, in *Proceedings of the CELE-Twente Workshop on Language Technology: Interacting Agents*, 71–82.

7. Qvortrup, L., ed. (2000) *Inhabited 3D Virtual Spaces*, Springer Verlag.
8. Traum, D.R. and Heeman, P.A. (1996) Utterance units and grounding in spoken dialogue, in *Proceedings of ICSLP*.
9. Core, M.G. and Allen, J.F. Allen. (1997) Coding dialogues with the DAMSL annotation scheme, in *Working Notes of the AAAI Fall 1997 Symposium on Communicative Action in Humans and Machines*.
10. Jönsson, A. (1997) A model for habitable and efficient dialogue management for natural language interaction, in *Natural Language Engineering*, Vol.3, 103–122, September.