M. THEUNE, D. HEYLEN, A. NIJHOLT

GENERATING EMBODIED INFORMATION PRESENTATIONS

Abstract. The output modalities available for information presentation by embodied, human-like agents include both language and various nonverbal cues such as pointing and gesturing. These human, nonverbal modalities can be used to emphasize, extend or even replace the language output produced by the agent. To deal with the interdependence between language and nonverbal signals, their production processes should be integrated. In this chapter, we discuss the issues involved in extending a natural language generation system with the generation of nonverbal signals. We sketch a general architecture for embodied language generation, discussing the interaction between the production of nonverbal signals and language generation, and the different factors influencing the choice between the available modalities. As an example we describe the generation of route descriptions by an embodied agent in a 3D environment.

1. INTRODUCTION

In conversations between human speakers, speech is the main carrier of information, but nonverbal signals such as gestures and facial expressions also play an important role, providing additional information about the content and the structure of the discourse. In order to successfully engage in a natural interaction with a human user, an embodied conversational agent should be able to interpret the user's speech and nonverbal signals, and to respond with appropriate verbal and nonverbal behaviours of its own. We know from our research on interaction with embodied conversational agents (Nijholt and Heylen, 2002) that extending the agent's repertoire of nonverbal behaviours can improve the quality of the interaction (Heylen et al., to appear).

Embodied conversational agents should therefore not just communicate by words but also nonverbally. Just as in human-human communication, many channels can be used to send a whole range of signals. To give a few examples, facial expressions can be used to express interest or surprise and a whole range of emotional and conversational signals. Posture shifts may indicate, among other things, a readiness to speak. Gestures can be used for several functions, such as accentuating parts of utterances. Gaze provides information about the focus of attention and plays a role in turn taking. Nonverbal signals can thus be used to signal information about the mental and emotional state of the agent, its personality, its understanding of what is said in the conversation and many other things.

In the past few years several groups have started to investigate different modes of nonverbal communication in more detail. This research deals with a whole range of questions: what signals are appropriate, what can we learn from human-human interaction in this respect, when do signals occur or when are they planned, how are

T.H.E. Editor(s) (ed.), Book title, 1—6.

they synchronized, how are they combined, how are they interpreted or evaluated by humans conversing with the agents, and so on.

The variety of channels and kinds of signals that can be examined, together with the large amount of parameters that enter into nonverbal communication means that much of the work is still exploratory, focussing on just a few behaviours and functions for specific settings. In principle one would like embodied agents to be capable not only of producing the appropriate signals, but also of receiving and interpreting the nonverbal signals of the human interlocutors. Although there are noticeable attempts at this (see e.g., Sowa et al., 2001, Breazeal, 2002, Wahlster, 2002), most work on nonverbal communication has been on the production side. We briefly discuss a few examples to illustrate the kinds of nonverbal signals that have been considered and the kind of research involved.

In the tutor agent Steve (Rickel and Johnson, 2000), nonverbal modes of communication have been introduced to help the instructions. Steve uses a number of signalling modes to aid in specific functions that are prominent in teaching situations, like drawing the attention of students to certain objects or actions and giving clues on how they are doing with their tasks. Body orientation, gaze and deictic gestures are used to direct the students' attention to objects in the virtual world. Steve can provide feedback to the students by shaking his head when telling students that they made an error or by a simple nod of approval when students perform correct actions. Nodding can also be used for back-channelling to acknowledge that Steve has understood the student. In this kind of research, the main task dictates the kind of use that is made of nonverbal communication.

Research in the Gestures and Narrative Language group of the MIT Media Lab, headed by Justine Cassell, has been concerned with studying aspects of nonverbal communication in human-human conversations and using these results to build computational models of that behaviour for conversational agents in specific dialogue settings. In these studies special attention is paid to the relation between the nonverbal acts on the one hand and linguistics aspects of the dialogue on the other hand. For instance, in their research on gaze, they looked at the role of gaze in relation to turn-taking and information structure. Other nonverbal behaviours they have investigated are posture shifts and beat gestures (Cassell et al., 2000ab, 2001a).

In "face-to-face" conversations, the face is the most expressive part of the body. Ekman (1979) provides an overview of the kind of signals that involve eyebrows. They play a role in common emotional expressions of sadness, surprise, fear and anger or distress; they may help underline certain words or punctuate the discourse; they may function as question markers, indicators of word search, et cetera. People are particularly sensitive to the eyes of others as a source of information about another person's mental state. Baron -Cohen (1995) speaks about a special language of the eyes. Facial expressions can be classified as either functioning as a conversational signal or as emotional icons. The latter could be expressive of an emotional state experienced by the conversant, but in face-to-face conversation, affective displays mostly have a social, communicative function (Kraut and Johnston, 1979; Chovil, 1992). Despite this, much work on facial expressions is solely concerned with a few emotional expressions. A notable exception is the work by Takeuchi and Nagao (1993), who pay attention to conversational signals as well.

A few other authors have looked in more detail at how to link facial expressions to the conversational actions and the intentional state of the agents (Pelachaud et al., 1996, 2002; Poggi et al., 2000).

The diversity in nonverbal communication is huge and involves many parameters, while our models of emotion, personality, dialogue and social interaction are still rudimentary. Given this, it is a daunting task to let embodied conversational agents converse naturally with the full repertoire of verbal and nonverbal modes of communication that people use in face-to-face conversations.

In this paper we will look at the relation between verbal and nonverbal communication for embodied conversational agents. We argue that because of their mutual dependence, the generation of verbal and nonverbal signals should be closely integrated. We will describe how a standard architecture for language generation can be extended to deal with the specification of nonverbal behaviours, and work out an example that shows how the resulting architecture can be used in a specific situation. First, however, we describe in more detail the role of human nonverbal modalities in conversation and discuss how taking these modalities into account can improve the interaction with conversational agents.

2. NONVERBAL SIGNALS IN HUMAN CONVERSATION

Humans engaged in conversation typically produce many different kinds of nonverbal signals, which include visual and audio signals. Gestures and facial expressions are the most noticeable visual signals; others are gaze and posture. Audio signals include non-speech sounds, like snorting and sighing, and prosody, which involves features of the speech signal like volume, tempo, and pitch. Both speakers and listeners produce such nonverbal signals, but with differences in frequency and type. Most nonverbal signals are produced unintentionally. In the case of gestures, it has been argued that their main function is to help the speaker with verbal formulation (Krauss et al., 1991; Rimé and Schiaratura, 1991). This point of view is supported by the fact that people still produce nonverbal signals when they cannot be seen, e.g., on the telephone. Still, even though many nonverbal signals may not be intended for the benefit of the hearer/viewer, it has been shown experimentally that people do make use of the information conveyed by such signals (see Kendon (1994) for an overview). In the following, we discuss three general classes of nonverbal signals, according to the kind of information they convey.

Following Pelachaud et al. (1996) and Cassell et al. (2000a), we distinguish between the interpretation of a nonverbal signal (its function, in terms of Cassell et al., 2000a) and its form: nonverbal signals with different forms can have the same interpretation, and the same nonverbal behaviour can be interpreted differently, depending on its context. The next three sections provide an overview of nonverbal signals in interaction between humans, grouped according to the kind of information they express. The last section discusses their relevance for embodied agents.

2.1 Expressing message content

One of the functions of communication is transferring knowledge: information about the state of the world is sent from the speaker to the listener. The information being transferred can be expressed using language, but also using nonverbal signals.

All nonverbal signals in this category are in principle ‘interchangeable’ with the corresponding verbal expressions. In normal circumstances, the main part of any message is expressed using language. In specific circumstances, ‘content-bearing’ signals can fully replace language. (Think of communicating with someone when there is too much noise to be heard!) Usually, however, they are used in combination with language, sometimes redundantly, and sometimes non-redundantly. With respect to timing, redundant signals are synchronized with the corresponding verbal expression. In the case of non-redundant signals, they co-occur with the verbal reference to the action, object or property that they provide additional information about. Nonverbal signals that are used to express message content can be divided into the following two classes.

Deictic signals are used to identify objects being referred to in the message. These signals usually take the form of a pointing hand gesture, but gaze direction, head and body movements are used as well. The indicated objects do not necessarily need to be visible, or even concrete. Deictic signals are different from other signals that express message content in that they do not inherently represent any meaning; their interpretation is entirely determined by the situational context. In this respect they are like verbal deictic expressions such as ‘here’, ‘we’, ‘now’, et cetera.

Representational signals are used to express concepts such as attributes, actions, and relationships between objects. They can be divided into different subclasses. Emblems are like words in that the relation between their shape and meaning is arbitrary and different across cultures. They can replace words or entire messages. Examples are nodding or head shaking to express agreement or disagreement, and the ‘thumbs up’ gesture for OK. Iconic and metaphoric signals illustrate properties of objects and actions. Their shape is not arbitrary, but reflects the meaning being expressed. In the case of iconic signals, there is a direct resemblance between the signal and the concept being depicted, for example hands forming a circular shape. They usually take the form of gestures, but facial expressions can also act as icons. For instance, squeezed eyes can symbolize small size (Poggi et al., 2000), or bad taste can be expressed by pulling a disgusted face. For pantomimes, the relation between shape and meaning is even more direct, whereas for metaphoric gestures some visual metaphor is used, such as depicting a physical container to represent a bearer of information, e.g., a film (McNeill, 1992).

2.2 Reflecting discourse structure

Nonverbal signals that reflect the structure of the ongoing discourse deal with the form rather than the content of the message. For instance, speakers mark focused discourse elements by intonation, quick hand movements (beat gestures), nods, eyebrow raises or combinations of these. Turn taking is associated with a number of nonverbal signals. Avoiding eye contact while speaking can indicate a reluctance to give away the turn whereas making eye contact can indicate a readiness to hand it over. Posture shifts, like leaning forward, are often used to indicate one wants to take the turn. Nonverbal signals like prosody and facial expressions can also mark the kind of speech act that is performed, like the typical prosody associated with questions, or back-channelling vocalisations to acknowledge what has been said.

These signals have well-defined relations with the linguistic channel, since what they do is provide information about the verbal discourse. In many cases, the information they convey can also be expressed verbally, but those expressions are often considered as marked. Examples are the use of cleft sentences ('It was John who…') and explicit turn taking phrases ('Do you have an ything to add?').

2.3 Showing speaker/hearer characteristics

Some signals do not convey information about the message, but about the person who is sending or receiving it. Face, gestures and voice may convey information about static characteristics such as personality, age, and gender, and about dynamic characteristics, holding at the moment of speaking: emotion (pleased, irritated), mental state (nervous, confused, paying attention or not), physical state (sick, sleepy, full of energy). This kind of information can be inferred from all kinds of nonverbal signals. In a few cases, the signals 'stand on their own' (most notably, facial expressions), but in most cases the information is read from the way other signals are performed, or from their frequency. For instance, a happy or extravert speaker will use broader gestures and more 'open' body postures; and if a speaker uses few gestures this can indicate tiredness or lack of enthusiasm, but also age or personality.

For most of these signals there is no direct temporal relation with the language being produced, although some dynamic states may be triggered by the content or form of the message (e.g., sadness at bad news, or irritation with long-windd wording). Speaker characteristics that are commonly expressed by nonverbal signals can be described verbally as well, e.g., the speaker can tell the listener about his current emotions. In this case, the information is part of the message content that is expressed intentionally (and which may be inaccurate). Usually, however, nonverbal signals of this type are produced unintentionally, and they are typically those that people will try to conceal or feign in face-to-face conversations (Ekman, 1992).

2.4 Relevance for Embodied Conversational Agents

We have seen that the production of nonverbal signals is an inherent part of human communication. Since nonverbal signals produced by human speakers can help the receiver to understand and remember the information being presented (cf. Kendon, 1994), this can be expected to hold for signals produced by embodied conversational agents as well. On the other hand, it has also been argued that nonverbal signals are usually not attended to, and that when they are, they are distracting and harm the processing of the message. Nevertheless, we believe that to allow for more fluent and natural interactions, embodied agents should be able to produce the same kinds of nonverbal signals as human speakers. Even if they are not always helpful, such nonverbal signals will make the embodied conversational agent more believable and lifelike, i.e., more like a human. This expectation is confirmed by our experiments on gaze (Heylen et al., to appear). For an overview of other experimental results on the effect of nonverbal signals (and agent embodiment in general) on human-agent interaction, see Dehn and van Mulken (2000).

3. AN ARCHITECTURE FOR EMBODIED LANGUAGE GENERATION

In this section we look at the generation of language in combination with nonverbal signals for the presentation of information by an embodied agent. Most existing embodied conversational agents or virtual presenters produce language using canned utterances. However, if the information that has to be presented changes over time, or is highly variable depending on the user's information nee ds or other contextual dynamics, this approach is not feasible, and some form of natural language generation is required. So far, research in natural language generation has been aimed primarily at unimodal information presentation, where some underlying message is expressed using natural language only, usually in the form of a written text. However, when the message is to be expressed by an embodied agent an additional modality becomes available in the form of nonverbal behaviour. This raises the question how language generation and the production of nonverbal signals should be combined.

A common approach is to specify the agent's verbal utterance first, and then add mark-up to indicate any accompanying nonverbal behaviour. In this approach, the language used by the agent is not adapted to its nonverbal actions. This may result in information presentations that are less concise than they could have been (since the nonverbal signals that are added can only reflect, but not complement or replace language), and also less natural (for instance, when a full description of an object is made superfluous by pointing). In section 2, we saw that most nonverbal signals have a strong relation with language. Therefore, our approach is to integrate the specification of nonverbal signals with language generation.

We focus on those nonverbal signals that are most closely related to language, i.e., deictic and representational signals, which mostly take the form of hand and arm gestures. The specification of other signals in connection to language generation is also discussed, but less extensively. In addition, we take a simplified view on information presentation as purely a 'monologue' task, and ignore nonverbal signals that are specific to dialogues, such as those signalling turn taking or dialogue acts. In the following, we first sketch a global architecture for generating embodied information presentations. Then, we discuss in some detail how the production of various kinds of nonverbal signals interacts with related language generation tasks.

3.1 General architecture

A standard architecture for language generation systems is described in Reiter and Dale (2000), who decompose the generation process into the following stages:

- *Document planning*. This involves determining the content and the global structure of the message to be presented. The outcome is an abstract message specification. In dialogue systems the dialogue planner generally carries out these tasks, which are largely domain specific and language independent.
- *Microplanning (or sentence planning)*. At this stage, the message specification is fleshed out further. This involves the generation of referring expressions, lexicalisation (word choice), and aggregation (grouping information into clauses and sentences). These tasks require both linguistic and domain knowledge.
- *Realisation*. Here, the abstract message specification is converted into real text, using knowledge about syntax, morphology, etc. In addition, mark-up may be added for use by external components.

We now discuss how the specification of nonverbal signals can be integrated in this general architecture, so that it can be used to generate texts that contain mark-up for nonverbal signals. In such an integrated architecture, the message specification created at the document planning (or rather, content determination) stage is the basis for specifying both verbal and nonverbal signals.

The two language generation tasks that are most closely related to the production of nonverbal signals are lexicalisation and the generation of referring expressions, both of which are carried out at the microplanning level. The main aim of referring expression generation is building a description that distinguishes the intended object from its distractors, i.e., other objects that might be referred to. Typically, such descriptions can be much reduced if they are accompanied by a deictic gesture that rules out most, or even all, distractors. Lexicalisation is the task of choosing words for expressing message concepts, such as actions and object properties. Since many concepts can be expressed by a representational nonverbal signal, the specification of such nonverbal signals can be seen as part of the lexicalisation process. The third microplanning task, aggregation, is related to the generation of nonverbal signals in that it delimits the domain for gesture production: on average, human speakers produce one gesture per clause during information presentation (McNeill, 1992), and to create a natural impression, a virtual presenter should do the same.

In the field of language generation, there is no consensus about the order of the three tasks involved in microplanning. Here, we make the following assumptions on this ordering. Referring expression generation comes before lexicalisation, since it may enrich the message specification with additional concepts (i.e., object properties) that must be lexicalised. For instance, to distinguish an object from its distractors it may be necessary to mention a certain property, e.g., its shape. This property will then have to be lexicalised, either verbally or by making an appropriate iconic gesture, or both. Aggregation is sometimes done before and sometimes after lexicalisation. As argued by Reiter and Dale (2000), it should come at least before referring expression generation, to avoid generating a description more often than necessary. For instance, the two messages *X is on R* and *Y is on R* can be aggregated to *X and Y are on R*, so that a reference to R needs to be generated only once. Given that referring expression generation should come before lexicalisation, this means that aggregation is the first microplanning task.

Nonverbal signals that reflect discourse structure, including prosody, do not really interact with language generation; rather, they reflect the structure of the result. Such signals are added at the realisation level, after syntactic realisation has been finished. The selection, both in form and frequency, of these and the other types of nonverbal signals should be influenced by different factors, such as the personality type or emotion the agent should convey (see sections 2.3 and 3.5). Ideally, such factors should also influence the language used by the system, a matter that is ignored in most language generation work (some exceptions are Hovy, 1988, de Rosis and Grasso, 2000, and, for embodied agents, André and Rist, 2000).

The output from embodied language generation is annotated text, specifying the nonverbal behaviour that has to be produced in parallel with certain words in the text. Several (XML-based) annotation languages have been proposed for this; e.g., Arafa et al., 2002, DeCarolis et al., 2002, Kranstedt et al., 2002. The annotated text is sent to external speech synthesis and animation modules. Speech synthesis converts the text into a speech signal, taking the prosodic markers into account. Animation takes care of producing the visual nonverbal signals, which have to be synchronised with respect to the pronunciation of the corresponding words, based on timing information from speech synthesis. For descriptions of how this can be done, see Pelachaud et al. (1996) and Cassell et al. (2001b). The animation component also has to determine the actual shape of the signals to be produced, taking relevant speaker characteristics into account (cf. Badler et al., 2002).

Figure 1 shows the general architecture for embodied language generation proposed here. We assume that it is a pipeline, i.e., there is no backtracking between modules. This means that nonverbal signals, once they have been added, cannot be removed, and that subsequent generation stages can only add nonverbal signals that do not conflict with those already specified. In other words, deictic signals take precedence over representational ones, and the latter over signals that reflect discourse structure. (We are aware that this is a simplification.) Figure 1 also shows the knowledge sources needed: discourse, domain, speaker, and user models, as well as world knowledge. At least some of these models must be dynamically updated during generation. In the following sections we see how such knowledge is used during the intertwined specification of language and nonverbal signals.

3.2 Deictic signals and the generation of referring expressions

References to objects in the message specification are potential opportunities for deictic nonverbal signals by the agent. A first selection from these candidates will only retain references to objects that are visible to the user.[1] Checking this condition is not always trivial; for instance, in applications where the user can move around in a 3D virtual world, as described in section 4, it will require some computation using

[1] Actually, this constraint is too strong, because human speakers also use deictic references to objects that are not visible, but occupy an imaginary position projected in the space before the speaker (McNeill, 1992). However, for simplicity we will ignore this here.

data from the domain model (listing the positions of virtual objects) and the user model (listing at least the user's c urrent position and orientation). Since human speakers use deictic gestures and other nonverbal signals mainly to convey new information, another condition is that the object should be new, i.e., not previously mentioned, inferable, or otherwise familiar to the user. We also assume that deictic signals are appropriate for objects that are mentioned in a contrastive context (for ways to detect contrast using a discourse model, see Prevost, 1995, Theune, 2002). In short, first or contrastive references to visible objects will be accompanied by a deictic nonverbal signal. We now discuss how this influences the creation of a verbal reference to the object.

Constructing a verbal description of an object, with the aim of allowing the user to uniquely identify it, involves selecting those properties of the object that distinguish it from other potential referents, the distractors. Having the agent point at a virtual object will generally rule out several of its distractors, namely those that are positioned outside the 'range' of the pointing finger. (We assume that pointing with the hand or finger is the preferred form of a deictic gesture.) When pointing is quite exact, for instance when the object is nearly touched, or when there are no distractors close to it, all distractors may be ruled out by the pointing gesture, so that no further verbal description is required. In such a situation, generating a syntactic placeholder such as *that* or *this one* may be sufficient.

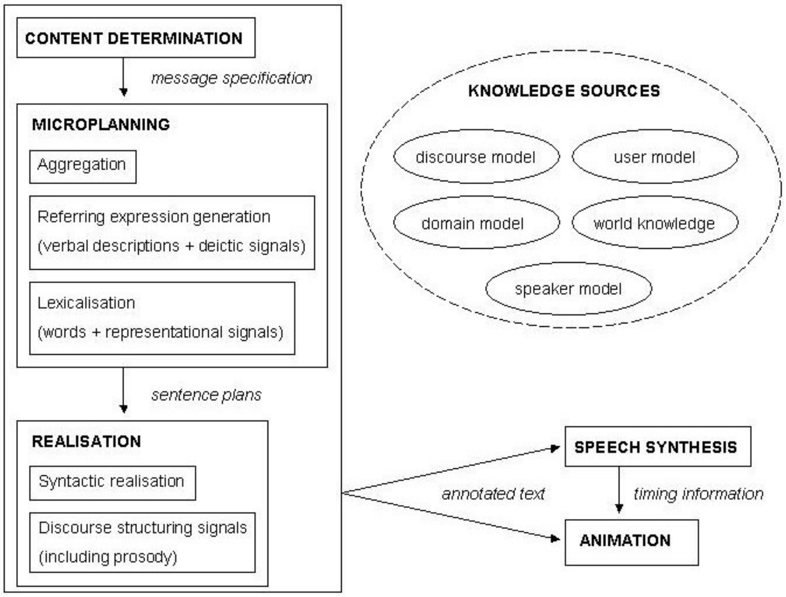

Figure 1. Architecture for combined generation of language and nonverbal signals.

In most cases, however, a pointing gesture by the agent will not rule out all distractors. As the distance from the agent to the virtual object increases, pointing becomes less precise (more or less in the fashion of a flashlight) and will rule out fewer distractors. In the case of deictic head movements or even only gaze, which must be reverted to if the agent's hands are occupied by some task, pointing is always very inaccurate, indicating only a general direction. So, depending on the distance from the agent to the intended referent, and the kind of deictic signal being produced, the system must determine which distractors are ruled out. The remaining distractors will have to be ruled out by the verbal description. The result of referring expression generation is a specification of the content of the verbal expression and of the deictic signal that should accompany it, if any. This information is added to the message specification, which is passed to lexicalisation.

3.3 Representational signals and lexicalisation

Not all concepts are equally suitable for expression using a representational nonverbal signal. Semantic features that can be easily visualized are the manner and direction of actions, and the shape, size and (relative) location of objects (Kendon, 1980, McNeill, 1992). When generating embodied information presentations, a simple way of checking whether a specific message concept can be expressed using a nonverbal signal is to use a 'gesture dictionary' or database linking concepts to nonverbal signals (Cassell et al., 2001b, Sowa et al., 2001). Such a dictionary may be based on a domain specific inventory of nonverbal signals that have been actually produced by human speakers. Nonverbal signals are only retrieved from the dictionary if they are compatible with nonverbal behaviours that were selected during the preceding generation task (i.e., deictic signals). For instance, the agent cannot produce an iconic gesture representing some property of an object if in the preceding stage it was decided that it should be pointing at that object. Similarly, representational head movements (e.g., a head wiggle to indicate dubiousness) are incompatible with deictic head movements. If multiple nonverbal signals are available to express a concept, a probabilistic choice can be made between them, based on their relative frequency in the corpus that was used to create the dictionary of representational signals.

Having identified those concepts in the message specification that the agent can express using a representational nonverbal signal, some of these must be selected. Simply grasping all opportunities for generating a nonverbal signal is not an option, as it is likely to produce an unnatural effect, e.g., giving the impression that the agent is "talking to a foreigner" (Cassell and Prevost, 1996). Like human speakers, in a neutral context the agent should not produce more than one or two gestures[2] per clause; a target that may be raised or lowered based on information from the speaker and user models. For instance, to convey an introvert personality, the agent should

[2] Most representational signals take the form of gestures; we have no figures available on the average number of other nonverbal behaviours produced by human speakers.

produce fewer nonverbal signals than average, whereas if the user actually *is* a foreigner, more nonverbal signals may be appropriate.

As with deictic signals, the selection of representational signals may be based on newness and contrast. Additional selection criteria may reflect the human tendency to express surprising or uncommon object features using a gesture, as well as information that is 'relevant to the primary communicative goal' (Yan, 2000). For cases where too many candidates meet these criteria, we need a ranking in the form of a (domain specific) information hierarchy, where the highest level is occupied by information that is most likely to be expressed using a nonverbal signal, e.g., new and highly relevant information, and where the lowest level is occupied by, e.g., discourse-old information. Based on the current target number of nonverbal expressions to be generated (see above), a selection can be made from the candidates, starting with those on the highest levels.

The next decision to be made is which concepts the agent should express only nonverbally, and which ones it should also express using speech, leading to a 'redundant' nonverbal signal (Cassell an d Prevost, 1996). Since nonverbal signals run a higher risk of being missed (i.e., overlooked) by the user than speech, it seems that the choice between redundant and non-redundant signals should be based on the deemed importance of the information to be expressed. In terms of the information hierarchy mentioned above, only items on the lower levels should be expressed non-redundantly. Other important factors are economy and ease of expression. Some concepts are more easily expressed nonverbally than verbally. For instance, complex motions or shapes can be relatively difficult or inefficient to express using a verbal description. This makes the use of speech less attractive, especially when the presentation is bound to certain time limits. Regardless of the information hierarchy, a non-redundant visual signal will be preferred in such cases.

In the choice between redundancy and non-redundancy, the kind of nonverbal signal being produced also plays a role. Emblems that are used non-redundantly seem to run little risk of being overlooked, because they actually replace the words they correspond with: speech is temporarily interrupted when a non-redundant emblem is produced, drawing the hearer's attention to the emblem. An example is saying *She is really <emblem for clever>*. Here, the choice for non-redundancy seems to be a matter of style rather than efficiency or effectiveness. On the other hand, other non-redundant representational signals are usually produced during the verbal description of the action or object of which they express an attribute. In this case, speech is not interrupted, so there is a risk of overlook, especially for subtle behaviours such as head or eye movements.

3.4 Nonverbal signals reflecting discourse structure

Nonverbal signals that reflect discourse structure take the form of both prosodic and visual nonverbal signals. Since these signals do not only mark informational but also syntactic structure, they are specified during the realisation stage, after syntactic realisation has taken place. Based on syntactic information,[3] some phrase boundaries can be marked prosodically by specific intonation contours or pauses of different lengths (cf. Theune et al., 2001), and visually by head motions, gaze direction, blinks or eyebrow movements (cf. Pelachaud et al., 1996). Similar visual signals, as well as beat gestures and pitch accents, can be used to mark words expressing new or contrastive information, and posture shifts may be inserted to mark topic changes (cf. Cassell et al., 2001a). As before, appropriate signals can only be selected if they are not in conflict with signals that were previously specified. In other words, we assume that both deictic and representational signals take precedence over signals that reflect discourse structure. This is similar to the rule of gesture class priority used in the BEAT system (Cassell et al., 2001b). Since deictic and representational signals most often take the form of gestures, in our proposed architecture discourse structure is somewhat more likely to be reflected by prosody and facial actions than by beat gestures, which run a higher chance of being incompatible with nonverbal signals specified during microplanning.[4] Rules for coordinating facial motions with intonation have been proposed by Pelachaud et al. (1996).

Finally, the influence of speaker characteristics must also be taken into account at this level. For instance, to create an impression of extraversion or enthusiasm, the agent should produce more beat gestures and pitch accents than average, and to convey introversion it should look away more and move less.

3.5 Factors influencing selection preferences and production rate

We have already discussed that the type and frequency of the nonverbal signals produced by the agent can be varied depending on which speaker characteristics should be conveyed, such as personality and emotion. These characteristics also influence the actual shape of the specified nonverbal signals as determined by the speech synthesis and animation modules. In addition, there are several other contextual factors that should be taken into account in the selection of nonverbal signals. Among these are user characteristics such as age, nationality, and level of expertise in the application domain. For instance, the use of more, less subtle nonverbal signals may be appropriate when the agent is speaking to a child as opposed to an adult, to a foreigner as opposed to a native speaker, and to an inexperienced user as opposed to an experienced one. In addition, there may be an influence of the application domain. In some domains, getting the message across correctly may be more important than in others; for example, in an educational setting production rates may be set higher than in social talk, and non-redundant signals may be avoided to ensure that the student does not miss any information. Within domains, message complexity may have to be taken into account. As shown by Cohen (1977) for the presentation of route descriptions, the gesture rate of human speakers increases with the complexity of the message being presented. Presumably,

[3] Theune et al. (2001) determine prosody on the basis of syntactic structure. On the other hand, Steedman (2000) points out that syntactic and prosodic structure may diverge.

[4] We ignore the possibility of overlaying beats over other gestures, as observed by McNeill (1992).

this also holds for other kinds of nonverbal signals and in other domains, and it may be sensible to have the agent copy such behaviour.

3.6 Discussion

In the preceding sections, we have taken a standard language generation architecture (Reiter and Dale, 2000) as the starting point for describing an 'ideal' system for the generation of embodied information presentations. However, several language generation systems do not adhere to this standard architecture, or do not perform all of the distinguished tasks (see Cahill et al., 1999). Also, in practice many generation tasks are carried out in an ad-hoc rather than a theoretically motivated manner. Examples are directly mapping concepts to standard lexicalisations instead of 'real' word choice, and the use of templates for syntactic realisation. Similarly, in the architecture sketched here all possible kinds of nonverbal signals are specified in a principled manner, but in practical systems, achieving this will be nearly impossible, and often unnecessary. Also, the model for embodied language generation presented here is necessarily simplified. Clearly, the interactions between language generation and the production of nonverbal signals are more intricate than has been presented here. Finally, much of what has been presented rests on assumption: there are many things about the production of nonverbal signals (and of language, for that matter) that we do not know yet. To extend our knowledge, careful study of human speaker behaviour will be required, as well as evaluation of implemented models. User experiments should be conducted to test the naturalness and effectiveness of different types of generated information presentations, as well as their effect on the user's per ception of the agent's personality and other speaker characteristics.

4. EXAMPLE: THE PRESENTATION OF ROUTE DESCRIPTIONS

In this section, we discuss the presentation of route descriptions as an example application of embodied language generation. Human speakers presenting a route description make prominent use of both deictic and representational signals to indicate landmarks and directions. These signals mainly take the form of broad hand and arm gestures; facial expressions play a relatively minor role. The ANGELICA[5] project aims at developing an embodied agent that presents route descriptions in the Virtual Music Centre (VMC, Figure 2). The VMC is a 3D virtual building with halls, corridors and different floors (Nijholt and Heylen, 2002). Since visitors to such environments often experience navigation problems (Nijholt et al., to appear), the VMC is a natural environment for the presentation of route descriptions. In previous research we developed a navigation agent that determines the user's intended destination within the VMC by means of a spoken dialogue, and then computes the shortest route to this destination (van Luin et al., 2001). This agent is not yet embodied, and cannot give a verbal route description: it either presents the

[5] A Natural-language Generator for Embodied, Lifelike Conversational Agents. This project is partially funded by NWO, the Dutch Organization for Scientific Research.

computed route on a map or moves the user through the VMC to the target location. In the ANGELICA project we will extend the navigation agent with a component for embodied language generation as sketched in the previous section. In the following, we describe how the embodied guide would generate a simple verbal and nonverbal route description, illustrating the proposed architecture.

Figure 2. Outside view of the VMC.

4.1 Content determination

The target location of the route description is determined in a spoken dialogue between the user and the embodied guide. This involves the generation of several dialogue utterances, accompanied by nonverbal signals. Several of these will be dialogue-specific, e.g., turn taking signals and signals that reflect different speech acts, such as confirmation and verification. Since our focus is on the presentation of the route description, we do not describe this stage of the interaction (but see Heylen et al. (to appear) for our work on gaze as a turn taking signal in spoken dialogues with an embodied agent). Here, we only assume that, after a number of turns, the target location is established as being the balcony of the VMC. The shortest route to this location is computed, and returned in the form of a vector of 3D coordinates. This vector is given as input to content determination, which turns it into a message specification as shown in Figure 3, which describes the route in terms of the actions the user has to carry out to reach the intended destination, for example walking towards some landmark or turning in some direction. For brevity, 'obvious' locations and actions, such as the start and end point of the route, have been left out. Creating such a route specification requires a nontrivial amount of domain and world knowledge, which is used to map the vector of 3D coordinates to actions, landmarks and spatial concepts. We do not go into that here.

```
ROUTE:[
  action:[type:walk,
          direction:up,
          object:s2],
  action:[type:turn,
          direction:sharp_right],
  action:[type:walk,
          direction:through,
          object:d3]]
```

Figure 3. Example route specification

4.2 Microplanning

The first task performed during microplanning is aggregation. Here, it is determined that the actions specified in our example message can be expressed in separate clauses, which are combined into one sentence. Within this sentence, the performer of each action (the user, left implicit in the message) remains the same, and therefore only needs to be mentioned in the first clause. In the following, we discuss the generation of referring expressions (including deictic signals) and lexicalisation (including representational signals) for each clause/action in turn.

Expressing the first action involves referring to the performer (the user) and the object of the action (s2, one of the two stairs in the VMC). For referring to the user, a simple pronoun ('you') is sufficient. This expression is not accompanied by a nonverbal signal, as the information being expressed is at the very bottom of the information hierarchy from section 3.3. The second reference is to the stairway, s2. This entity is new to the discourse, essential for the route description, and in the user's current line of si ght. A deictic hand gesture is therefore selected. To make clear that the agent refers to the stairs as a whole, and not to one of its steps or the handrail, the referring expression algorithm specifies that the object's type property must be added to the verbal part of the expression. In combination with the gesture, this is sufficient to uniquely identify object s2. (If pointing had not been possible, for instance because the user could not see s2, then a more elaborate verbal description distinguishing it from the other stairway in the VMC would have been required.)

Now that the content of the referring expressions is known, lexicalisation starts. One candidate for nonverbal expression is s2's property of being a stairway, which can be easily expressed using, for instance, an iconic gesture representing the shape of the steps (assuming such a gesture is present in the 'gesture dictionary'). However, a deictic gesture has already been selected for referring to the stairs, leaving no room for any additional hand movements by the agent. Therefore, the type property can only be lexicalised verbally. Another candidate for nonverbal expression is the upward direction of the action. This information is both new and essential for the route to be taken, so an upward moving gesture is selected. The action type, walking, might be expressed by a 'walking finger' gesture (cf. Cassell et al., 2000a). However, this information is low in the information hierarchy, since there is no other way to move around in the VMC. Therefore no nonverbal expression is selected to express this aspect of the action, and on the verbal side the general verb 'go' is preferred over the more specific, but superfluous, 'walk'. All in all, two gestures are selected for the first clause: pointing upwards and at the stairs.

Describing the second action does not involve references to objects, and thus no deictic signals. During lexicalisation, it is found that the concept of turning in a specific direction can be expressed nonverbally. Since this concept is new and important, a gesture is selected where the hand and arm demonstrate a sharp right turn. On the verbal side, the phrase 'turn sharply to the right' is selected. Here, the adverbial 'sharply' might have been left out, as it is grammatically optional and the sharpness of the turn is already illustrated by the gesture. However, this aspect is essential for the route, since the user can also make a non-sharp right turn at the top of the stairs. Only using a non-redundant gesture carries the risk of this information being overlooked and the user taking a wrong turn, so an additional verbal expression is preferred.

The third action involves d3, one of the doors in the VMC. Because d3 is invisible from the user's current location a deictic signal is deemed inappropriate, and a distinguishing verbal description must be created. The VMC has several doors, but only one of them (d4) is near to d3 and therefore counts as a distractor. To distinguish d3 from d4, describing it as 'the first door' suffices. We assume that the door's attribute of being first, or nearest, cannot be expressed nonverbally.[6] The concept of (walking through) a door may be expressed by a nonverbal signal, e.g., a pantomimic gesture. However, walking through is an obvious action in connection with a door, and so we assume that no such gesture will be selected and that this information is only expressed verbally.

4.3 Realisation

The final generation stage is realisation. For the first clause, syntactic realisation produces the sentence 'You go up the stairs'. Then, nonverbal signals reflecting discourse structure are added. The clause is too short for pauses, but the end of the clause is marked prosodically by rising intonation, indicating that the sentence is not yet at its end. The words 'up' and 'stairs' convey new information, so these receive a pitch accent. They cannot be accompanied by beat gestures, since the agent's hands are already involved with gestures specified during microplanning: pointing upwards and at the stairs. Available alternatives to mark the focused information are blinking and raising the eyebrows (cf. Pelachaud et al., 1996).

The second clause is syntactically realised as 'turn sharply to the right', where 'sharply' and 'right' are accented and accompanied by blinks and eyebrow movements. Again, there is no room left for beat gestures. The final clause is realised as 'and go through the first door'. Here, 'first' is accented and accompanied by a beat gesture, since this word expresses information that is both new and

[6] A representational gesture such as sticking up an index finger or thumb to represent the concept one ('first') does not seem quite appropriate in this case, as it would rather be used for indicating a number of objects (*One beer please*) or when summing up several points (*First, …; second, …*). However, determining this may require a more sophisticated approach than simple dictionary look-up.

contrastive (distinguishing d3 from d4). The word ‘door’, on the other hand, is not accented, since this property is shared by d3 and d4 and thus not contrastive. The end of the clause is marked by a falling intonation contour, since it is sentence-final. The resulting route description, with simplified mark-up, looks as follows:

> you go **<g type=iconic>**UP**</g>** the **<g type=deictic>**STAIRS**</g>** //
> **<g type=iconic>**turn SHARPLY to the RIGHT**</g>** //
> and go through the **<g type=beat>**FIRST**</g>** door ///

Here, accented words are given in small capital letters, and phrase boundaries of different strengths are indicated by a number of slashes. Of the visual signals, only gestures are indicated using greatly simplified XML style mark-up. The embodied presentation of the first clause is illustrated in Figure 4.

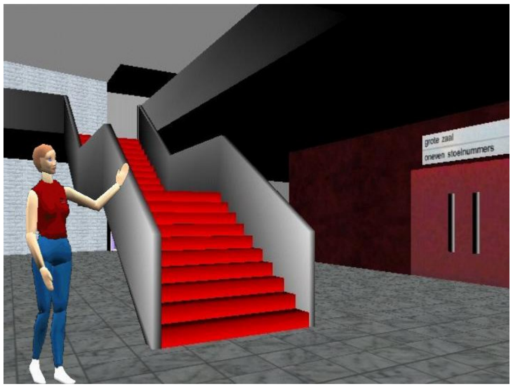

Figure 4. ‘You go UP the STAIRS’

4.4 Remaining issues

It is clear that generating an embodied route description does not only require rules for the generation of verbal and nonverbal signals, but also involves a good deal of reasoning on the basis of (spatial) domain knowledge and general knowledge. The questions that need to be resolved include, which information is more or less essential for describing the route, how can ‘virtual objects’ and their properties be linked to coordinates in the 3D world, which objects are currently visible to the user (or will be from certain points along the route), and which objects are within the range of a deictic signal. In addition, it is not clear how to react to user interruptions or movements, during information presentation. For instance, what if the user moves away from the virtual guide, perhaps to look where she is pointing?

Other remaining problems are at the level of speech synthesis and, in particular, animation. One of these is timing: the generated nonverbal signals have to be synchronized with the pronunciation of the corresponding words. For signals that correspond to phrases of varying lengths, this might require speeding up or slowing down their execution, depending on speech times. If we assume that the guide does not have a fixed location, this means that deictic signals must be animated in real time, based on the current position of the guide relative to the object being indicated. This involves selecting from all possible joint rotations and translations those that will result in a natural looking movement. Another issue is that of realistically blending similar signals (e.g., different gestures or facial expressions) that are performed in close sequence. An example is combining the upwards and deictic gestures from the first clause of our route description. Nonverbal facial signals will also have to be combined with facial movements that are directly related to physical speech production (visemes). Finally, in determining the actual shape of the nonverbal signals, both in animation and in speech synthesis, any relevant speaker characteristics should be taken into account. For many of the research issues listed above, practical solutions are still lacking, so that at the implementation level several of these problems will have to be sidestepped or handled in an ad-hoc manner.

5. RELATED WORK

So far, there has been relatively little work on combining language generation with the production of nonverbal signals by embodied conversational agents. Notable exceptions are the work by Cassell et al. (2000ab) and Pelachaud et al. (2002). Cassell et al. focus on the division of labour between speech and representational gestures, guided by information about the discourse context and the communicative function of the utterance to be generated. Their work is applied in REA, a virtual real estate agent that presents embodied descriptions of houses. In addition to representational signals, REA also generates nonverbal signals that are related to discourse structure. These include posture shifts (Cassell et al., 2001a), gaze, and beat gestures. Pelachaud et al. (2002) developed the Greta agent, a talking face that presents medical diagnoses to patients. Greta can display nonverbal signals that reflect discourse structure and the agent's cognitive and emotional state, as well as signals that express message content. There is no interaction between language generation and nonverbal signal production, however. Other research in this area is that of André and Rist (2000), who have worked on language generation for virtual presenters, focusing on the problem of projecting different agent personalities. The NECA project (Krenn et al., 2002) builds further on this work, developing components for reasoning about agents' emotions and generating combined (emotional) speech and nonverbal signals. Work on discourse-related prosody combined with language generation includes that of Williams (1999) (in the route description domain), McKeown and Pan (2000) and Theune et al. (2001).

Schmauks (1987), Claassen (1992), Reithinger (1992), and Lester et al. (1999) address the generation of multimodal referring expressions. Their approaches are not aimed at animated agents, with the exception of Lester et al., who combine language generation with the production of pointing gestures by a pedagogical agent. All mentioned approaches are limited in that pointing is assumed to be exact, and hardly influences property selection for the verbal description. Krahmer and van der Sluis (2003) propose a new model for generating multimodal references that improves on this by using various degrees of pointing precision, and relating the selection of distinguishing properties to the kind of pointing gesture being generated.

The presentation of route descriptions by an embodied agent is addressed in the REAL project, where an embodied agent shows users the way through a virtual environment (Baus et al., 2000). However, this agent has no language generation capabilities except for limited object references. In the MACK system (Cassell et al., 2002), route descriptions are presented by an embodied conversational agent that is projected in a real-world environment. This agent does not have language generation capabilities, but uses canned text. It produces nonverbal signals using a probabilistic strategy, based on data from route descriptions by human speakers.

6. CONCLUSION

Over the years there has been a steady increase of our knowledge of human verbal and nonverbal communication, and how it can be simulated by embodied conversational agents. Nevertheless, the actual development of applications in which all kinds of contextually appropriate nonverbal signals are generated in a principled fashion is still some steps away. In this chapter we have sketched a general architecture for combining language generation with the specification of nonverbal signals for information presentation by embodied conversational agents, and illustrated how it can be used in a specific application. The model for embodied language generation we have presented here is necessarily simplified, and much of it rests on assumption. To extend our knowledge, further studies of spontaneous human speaker behaviour are required, as well as controlled experiments and evaluations of implemented systems. Our ultimate goal is to allow for human-agent interactions that are equally natural, effective, and entertaining as face-to-face interactions between humans.

REFERENCES

E. André and T. Rist (2000). Presenting through performing: on the use of multiple lifelike characters in knowledge-based presentation systems. *Proceedings of the Second International Conference on Intelligent User Interfaces*, New Orleans, USA, 1-8.

Y. Arafa, K. Kamyab, E. Mamdani, S. Kshirsagar, N. Magnenat-Thalmann, A. Guye-Vuilleme and D. Thalmann (2002). Two approaches to scripting character animation. *Proceedings of ECA' s: Let' s specify and evaluate them! Workshop held in conjunction with AAMAS 2002*.

N. Badler, J. Allbeck, L. Zhao and M. Byun (2002). Representing and parameterizing agent behaviors. *Proceedings of Computer Animation, IEEE Computer Society,* Geneva, Switzerland, 133-143.

S. Baron-Cohen (1995). *Mindblindness*. MIT Press.

J. Baus, A. Butz and A. Krüger (2000). Incorporating a virtual presenter in a resource adaptive navigational help system. *Proceedings of the Workshop on Guiding Users through Interactive Experiences*, Paderborn, Germany.

C.L. Breazeal (2002). *Designing Sociable Robots*. MIT Press.

L. Cahill, C. Doran, R. Evans, C. Mellish, D. Paiva, M. Reape, D. Scott and N. Tipper (1999). In search of a reference architecture for NLG systems. *Proceedings of the 7th European Workshop on Natural Language Generation (EWNLG' 99,)*Toulouse, France, 77-85.

J. Cassell and S. Prevost (1996). Distribution of semantic features across speech and gesture by humans and computers. *Proceedings of the Workshop on the Integration of Gesture in Language and Speech*, Newark, USA, 253-270.

J. Cassell, T. Bickmore, L. Campbell, H. Vilhjálmsson and H. Yan (2000a). Conversation as a system framework: Designing embodied conversational agents. In J. Cassell, S. Prevost, E. Churchill, and J. Sullivan (eds.), *Embodied Conversational Agents*. MIT Press, 29-63.

J. Cassell, M. Stone and H. Yan (2000b). Coordination and context-dependence in the generation of embodied conversation. *Proceedings of the First International Conference on Natural Language Generation*, Mitzpe Ramon, Israel, 171-178.

J. Cassell,Y. Nakano, T. Bickmore, C. Sidner and C. Rich (2001a). Non-verbal cues for discourse structure. *Proceedings of the 41st Annual Meeting of the Association of Computational Linguistics (ACL 2001)*, Toulouse, France, 106-115.

J. Cassell, H. Viljhálmsson and T. Bickmore (2001b). BEAT: the Behavior Expression Animation Toolkit.*Proceedings of the 28th Int. Conference on Computer Graphics and Interactive Techniques (SIGGRAPH 2001),* Los Angeles, USA, 477-486.

J. Cassell, T. Stocky, T. Bickmore, Y. Gao, Y. Nakano, K. Ryokai, D. Tversky, C. Vaucelle and H. Vilhjálmsson (2002). MACK: Media lab Autonomous Conversational Kiosk. *Proceedings of Imagina ' 02*Monte Carlo, Monaco.

N. Chovil (1991/1992). Discourse-oriented facial displays in conversation. *Research on Language and Social Interaction 25*, 163-194.

W. Claassen (1992). Generating referring expressions in a multimodal environment. In R. Dale, E. Hovy, D. Rösner and O. Stock (eds.), *Aspects of Automated Natural language Generation*. Springer Verlag, Berlin, 247-262.

A. Cohen (1977). The communicative functions of hand illustrators. *Journal of Communication 27*, 54-63.

A. Cohen (1980). The use of hand illustrators in direction-giving situations. In W. von Raffler-Engel (ed.), *Aspects of Nonverbal Communication*, Swets and Zeitlinger BV, Lisse, 265-273.

B. DeCarolis, V. Carofiglio, M. Bilvi and C. Pelachaud. APML, a mark-up language for believable behavior generation. *Proceedings of ECA' s: Let' s specify and evaluate them! Workshop held in conjunction with AAMAS 2002*, Bologna, Italy.

D. Dehn and S. van Mulken (2000). The impact of animated interface research: a review of empirical research. *International Journal of Human-Computer Studies 52(1),* 1-22.

P. Ekman (1979). About brows. In M. von Cranach, K. Foppa, W. Lepenies and D. Ploog (eds.) *Human Ethology*, Cambridge University Press, 169-202.

P. Ekman (1992). *Telling Lies*. W.W. Norton and Company.

D. Heylen, I. van Es, A. Nijholt and B. van Dijk (to appear). Controlling the gaze of a conversational agent. In J. van Kuppevelt, L. Dybkjaer and O. Bernsen, *Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems*. Kluwer Academic Publishers.

E. Hovy (1988). *Generating Natural Language Under Pragmatic Constraints*. Lawrence Erlbaum, Hillsdale, New Jersey.

A. Kendon (1980). Gesticulation and speech. In M.R. Key (ed), *The Relationship of Verbal and Nonverbal Communication*. Mouton, Den Haag, 207-227.

A. Kendon (1994). Do gestures communicate? A review. *Research on Language and Social Interaction 27(3),* 175-200.

E. Krahmer and I. van der Sluis (2003). A new model for the generation of multimodal referring expressions. *Proceedings of the 9th European Workshop on Natural Language Generation, held in conjunction with EACL2003,* Budapest, Hungary, 47-54.

A. Kranstedt, S. Kopp and I. Wachsmuth (2002). MURML: a Multimodal Utterance Representation Markup Language for conversational agents. *Proceedings of ECA' s: Let' s specify and evaluate them! Workshop held in conjunction with AAMAS 2002*, Bologna, Italy.

R.M. Krauss, P. Morrel-Samuels and C. Colasante (1991). Do conversational hand gestures communicate? *Journal of Personality and Social Psychology 28*, 389-450.

R.E. Kraut and R.E. Johnston (1979). Social and emotional messages of smiling: an ethological approach. *Journal of Personality and Social Psychology 37*, 1539-1553.

B. Krenn, H. Pirker, M. Grice, S. Baumann, P. Piwek, K. van Deemter, M. Schroeder, M. Klesen and E. Gstrein (2002). Generation of multimodal dialogue for net environments. In S. Busemann (ed.), *Konferenz zur Verarbeitung natürlicher Sprache (KONVENS 2002)*, Saarbrücken, Germany, 91-98.

J. Lester, J. Voerman, S. Towns and C. Callaway (1999). Deictic believability: coordinating gesture, locomotion, and speech in lifelike pedagogical agents. *Applied Artificial Intelligence 13 (4-5)*, 383-414.

J. van Luin, A. Nijholt and R. op den Akker (2001). Natural language navigation support in virtual reality. *Proceedings of the International Conference on Augmented, Virtual Environments and Three-dimensional Imaging (ICAV3D).* Mykonos, Greece, 263-266.

K. McKeown and S. Pan (2000) Prosody modeling in concept-to-speech generation: methodological issues*, Philosophical Transactions of the Royal Society 358,* 1419-1431.

D. McNeill (1992). *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press, Chicago.

A. Nijholt and D. Heylen (2002). Multimodal communication in inhabited virtual environments. *International Journal of Speech Technology 5*, 343-354.

A. Nijholt, J. Zwiers and B. van Dijk (to appear). Maps, agents and dialogue for exploring a virtual world. In J. Aguilar, N. Callaos and E.L. Leiss (eds.), *Web Computing*. International Institute of Informatics and Systemics (IIIS).

C. Pelachaud, N. Badler and M. Steedman (1996). Generating facial expressions for speech. *Cognitive Science 20*, 1-46.

C Pelachaud, V. Carofiglio, B. De Carolis, F. de Rosis and I. Poggi (2002). Embodied contextual agent in information delivering application. *Proceedings of the First International Joint Conference on Autonomous Agents & Multi-Agent Systems (AAMAS' 02)*Bologna, Italy, 758-765.

I. Poggi, C. Pelachaud, and F. DeRosis (2000). Eye communication in a conversational 3D synthetic agent. *AI communications 13(3),* 169-182.

S. Prevost (1995). *A Semantics of Contrast and Information Structure for Specifying Intonation in Spoken Language Generation*. PhD thesis, University of Pennsylvania.

E. Reiter and R. Dale (2000). *Building applied natural language generation systems*. Cambridge University Press, Cambridge.

N. Reithinger (1992). The performance of an incremental generation component for multimodal dialog contributions. In R. Dale, E. Hovy, D. Rösner and O. Stock (eds.), *Aspects of Automated Natural language Generation*. Springer Verlag, Berlin, 263-276.

J. Rickel and W. L. Johnson (2000). Task-oriented collaboration with embodied agents in virtual worlds. In J. Cassell, S. Prevost, E. Churchill and J. Sullivan (eds.), *Embodied Conversational Agents*. MIT Press, 95-122.

B. Rimé and L. Schiaratura (1991). Gesture and speech. In R. Feldman and B. Rimé (eds.), *Fundamentals of Nonverbal Behavior*. Cambridge University Press, Cambridge, 239-281.

F. de Rosis and F. Grasso (2000). Affective natural language generation. In A.M. Paiva (Ed.), *Affective Interactions*. Springer Lecture Notes in AI 1814, 204–218.

D. Schmauks (1987). Natural and simulated pointing. *Proceedings of the 3rd Conference of the European Chapter of the Association for Computational Linguistics (EACL' 87)*79-185.

T. Sowa, S. Kopp, M. E. Latoschik (2001). A communicative mediator in a virtual environment: processing of multimodal input and output. *Proceedings of the International Workshop on Information Presentation and Natural Multimodal Dialogue (IPNMD-2001)*,Verona, Italy, 71-74.

M. Steedman (2000). Information structure and the syntax-phonology interface. *Linguistic Inquiry 31(4),* 649-689.

A. Takeuchi and K. Nagao (1993). Communicative facial displays as a new conversational modality, *InterCHI' 93*Amsterdam, The Netherlands, 187-193.

M. Theune (2002). Contrast in concept-to-speech generation. *Computer Speech and Language 16(3/4)*, 491-531.

M. Theune, E. Klabbers, J.R. de Pijper, J. Odijk and E. Krahmer (2001). From data to speech: a general approach. *Natural Language Engineering 7(1),* 47-86.

W. Wahlster (2002). SmartKom: fusion and fission of speech, gestures, and facial expressions. *Proceedings of the First International Workshop on Man-Machine Symbiotic Systems*, Kyoto, Japan, 213-225.

S. Williams and C. Watson (1999). A profile of the discourse and intonational structures of route descriptions. *Sixth European Conference on Speech Communication and Technology (Eurospeech' 99*, Budapest, Hungary, 1659-1662.

H. Yan (2000). *Paired Speech and Gesture Generation in Embodied Conversational Agents*. Master's thesis, Media Lab, MIT.