Embodied Conversational Characters: Representation Formats for Multimodal Communicative Behaviours

Brigitte Krenn, Catherine Pelachaud, Hannes Pirker, and Christopher Peters

Abstract This contribution deals with the requirements on representation languages employed in planning and displaying communicative multimodal behaviour of embodied conversational agents (ECAs). We focus on the role of behaviour representation frameworks as part of the processing chain from intent planning to the planning and generation of multimodal communicative behaviours. On the one hand, the field is fragmented, with almost everybody working on ECAs developing their own tailor-made representations, which is amongst others reflected in the extensive references list. On the other hand, there are general aspects that need to be modelled in order to generate multimodal behaviour. Throughout the chapter, we take different perspectives on existing representation languages and outline the fundamental of a common framework.

1 Introduction

This contribution deals with the requirements on representation languages employed in planning and displaying communicative multimodal behaviour of embodied conversational agents (ECAs). The term ECA has been coined in Cassell et al. (2000) and refers to human-like virtual characters that typically engage in a face-to-face communication with the human user, employing various synchronised channels of communication such as facial expression, hand–arm gestures and body posture, as well as tone of voice and text. The embodiment of ECAs ranges from talking heads to full-bodied 2D and 3D characters. Underlying are complex AI systems that model the character's capabilities to meaningfully engage in communicative situations with the human user or other ECAs. This includes modelling of the character's self, its tasks, goals, related believes and intentions, its personality, emotions and interpersonal stances, but also modelling of the character's (perceived)

B. Krenn (✉)
Austrian Research Institute for Artificial Intelligence, Freyung 6, 1010, Vienna, Austria
e-mail: brigitte.krenn@ofai.at

P. Petta et al. (eds.), *Emotion-Oriented Systems*, Cognitive Technologies,
DOI 10.1007/978-3-642-15184-2_20, © Springer-Verlag Berlin Heidelberg 2011

environment, including social scenarios, communication situations and partners, in order to generate situationally adequate and believable behaviours.

ECA systems may implement a face-to-face dialogue with the user, model scenarios where humans and artificial agents interact with each other in a virtual or mixed environment or generate communicative interactions between different artificial characters in which the user actively takes part or which are displayed to the user like in a product presentation, a TV spot or a stage play. See for instance André and Rist (2000) and Nijholt (2006) where one or more virtual agents present information to the user, Krenn et al. (2002) where a virtual car seller and buyer engage in a conversation about various features of cars and Mateas and Stern (2003) where in an interactive drama the interaction of a user with two characters representing a couple on the verge of divorce influences the outcome of the couple's story. Rehm and André (2005) present GAMBL, an interactive test-bed for human–ECA interaction. Other examples of ECA implementations are the Real Estate Agent REA (Cassell et al., 1999) which implements a full perception–action loop of communication by interpreting multimodal user input and generating multimodal agent behaviour; the pedagogical agent Steve (Rickel and Johnson, 1998) which functions as a tutor in training situations; MAX (Kopp and Wachsmuth, 2004), a virtual character geared towards simulating multimodal behaviour; and Carmen (Marsella et al., 2003), a system that supports humans in emotionally critical situations such as advising parents of infant cancer patients. ECAs can adopt several roles, such as being a teacher (Johnson et al., 2005; Moreno, 2007), a doctor (De Rosis et al., 2003), a museum guide (Kopp and Wachsmuth, 2004), a real-estate agent REA (Cassell et al., 1999) or a companion (Bickmore and Cassell, 2005; Hall et al., 2006). Even though the above examples represent only a small fraction of the vast and constantly growing work on ECAs, they are well suited for illustrating the broad range of applications.

Human communicative behaviour covers a broad range of skills, including natural language generation and production, co-verbal gesture, eye gaze and facial expression. People produce such behaviours in real time with ease and in a broad range of circumstances. In order to simulate human-like multimodal communicative behaviour, advanced ECA systems need to incorporate a whole range of complex processing steps, from intent to behaviour planning to behaviour realisation including some sort of scene or story generation, multimodal natural language generation, speech synthesis, the temporal alignment of verbal and non-verbal behaviours and behaviour realisation employing particular animation libraries and engines. In the current contribution, we focus on the role of behaviour representation frameworks as part of the processing chain from intent planning to the planning and generation of multimodal communicative behaviours.

2 Background

Imagine a situation, where we want to model the following encounter between a character C and a user U: C is in a good mood, encountering the appearance of U in the system makes C particularly happy and leads C to greet U effusively. Finally, we want to see an animation including the following behaviours: C displays a neutral but friendly face, directs its attention to U, broadly smiles at U and says *Hello my friend! Good to see you after such a long time!* As regards the spoken utterance, we want to put emphasis on *hello, good* and *such*.

At the intentional level we thus have something like the following ingredients which we present in a pseudo-notation (in reality XML[1] formats are widely used):

Example 1. Intentional Level pseudo-notation

```
mood(C) = happy;
event = encounter(C,U) -->
emotion(C) = happy & communication_act(C) = greet(U).
```

At the behaviour planning level, we need to further specify the non-verbal and verbal behaviours, bring them into a temporal order and specify the relative dependencies between the communication channels involved. Employing our pseudo-notation, this might be represented as follows:

Example 2. Behaviour planning level pseudo-notation

```
c1: face(C) = neutral_friendly;
c2: gaze(target(C))= user;
c3: face(C) = broad_smile;
c4: gesture(C) = wave;
c5: utterance(C) = emph {hello} my friend! m1 emph{good} to
see you after such a emph{long} time!;
start(c1) = t0;
start(c2) = encounter(C,U);
start(c3) = encounter(C,U);
end (c3) = start(c5);
start(c4) = encounter(C,U);
end(c4) = m1.
```

The above notation defines the partial behaviours (c1 to c5) including face, gaze, gesture and utterance, puts them in a relative temporal order and specifies the following dependencies: At time t0, the beginning of the animation, C looks neutral_friendly and starts to look at the user when encountering them. At the same time, C starts a broad smile and waves. When beginning to speak C stops smiling. The wave ends after *friend* has been spoken out and before the onset of *good*.

At the behaviour realisation level, we need to specify the actual behaviours at an even greater level of detail such that the realisation engine is able to generate a sequence of integrated multimodal behaviours for playing. In particular, the concrete animations are selected, the utterance is synthesised and the timing of the partial behaviours is transformed from relative to absolute. For lip-synchronised

[1] http://www.w3.org/XML/

speech, phonemes (transcripts of consonants and vowels) are aligned with visemes (visualisations of mouth shapes which may also include the tongue). Depending on the speech synthesis component employed, also a markup of prosodic information including syllables, intonation phrases and related accents may be available. This information is necessary to synchronise eyebrow raises and accents.

As our small example has shown, there is a variety of information that needs to be modelled in order to represent multimodal communicative behaviour. Accordingly, considerable effort has been put into the development and documentation of representation formats in the last years. The overall complexity of ECA systems motivated different strands of development. On the one hand it gave rise to a number of XML-based markup languages which are aimed at providing means for human authors to easily annotate text with multimodal behaviours. On the other hand attention was geared towards the specification of representation formats for the exchange of information between sub-components of an ECA system.

The goal of this contribution is to give an overview on representation formats proposed so far and will discuss specific representational needs raised by selected sub-tasks such as modelling at different levels of processing emotional display and spoken dialogue accompanying bodily behaviours. An evaluation in terms of the general acceptance and dissemination of different representation formats will be provided. Moreover, we will present an initiative for a common architecture for ECA systems and the prospects for the future development of representation languages. The community is still investigating ways to come up with strategies of unifying the existing variety of representation formats. We expect benefits of such an endeavour only if representation formats and source codes of related processing components are made available for free to the research community.

3 Different Views on Representation Languages/Formats for Behaviour Generation of ECAs

Numerous representation languages or formats have been proposed in the literature. They include markup languages for annotating text with behaviour directives, representation languages that declaratively and to different degrees of detail model various aspects of information required at different stages of behaviour generation, and languages that incorporate procedural knowledge in their annotations. Some languages attempt to cover a broad range of information relevant for behaviour generation. Most of the representations, however, have been designed for specific applications. In order to structure the wealth of proposals, we will, in the following, offer two views on existing representations: First we provide examples for different representation formats ranging from text markup to representations that contain aspects of high-level programming languages. Second we will present examples for representation languages that attempt to cover a broad range of information versus representations that have been designed with a specific application purpose in mind.

3.1 Formats – Markup– Versus Representation– Versus Scripting Languages

Markup languages typically define sets of markups that give the non-expert user (usually a web designer) the possibility to annotate text with high-level behavioural information in order to produce pre-scripted presentations for ECAs. VHML (Beard and Reid, 2002)[2] is an example for this type of languages. It has been designed for creating interactive multimodal applications with talking heads or full-bodied ECAs. Other examples of ECA markup languages where text is annotated with high-level concepts are APML (De Carolis et al., 2004) and MPML (Zong et al., 2000).

Representation languages in contrast aim for the technically detailed annotation of theory-specific information. In this respect, the Emotion Annotation and Representation Language (EARL) addressed in Schröder et al. (2010) is more a representation than a markup language. This holds in general for all languages that become more and more detailed in modelling and describing multimodal behaviours. Thus representation languages are well suited to function as data representation formats inside a system, especially as representations at the interfaces between the individual sub-components. RRL (Rich Representation Language[3], Piwek et al., 2002) is an example for such a language that defines an XML format for representing the input and output of all the components used for realising the processing steps from intent to behaviour planning to behaviour realisation.

Scripting languages in addition also incorporate means for encoding procedural knowledge, e.g. conditional execution of behaviours such as “if event X occurs, then execute behaviour Y”. Thus scripting languages are comparable to high-level programming languages. Examples in the field of ECAs are STEP and its XML variant XSTEP (Huang et al., 2003), and ABL (Mateas and Stern, 2004). The expressive power of scripting languages comes with a price though, e.g. the complexity of writing specifications in ABL comes close to programming in Java. The choice of the appropriate level of representation thus has to take different constraints into account. On the one hand, markup languages are indispensable in application development, because the application designer need not necessarily be an expert in all the fields underlying the development of ECA systems. On the other hand, representation languages are crucial in research contexts, because of the necessity to represent highly specific, low-level information. With the increased demand for truly interactive systems, the need for including at least some procedural capabilities typical for scripting languages into the representations becomes more and more of an issue, resulting in new hybrid formats. This trend is, e.g., exemplified in the evolution from MPML to MPML3D (Nischt et al., 2006). While MPML was designed as a markup language that allows the non-expert user to create pre-synchronised presentations, MPML3D has evolved to an authoring system that allows for the embedding of scripts and for the design of reactive scenarios. Summing up, while multimodal

[2] http://www.vhml.org, Virtual Human Markup Language

[3] http://www.ofai.at/research/nlu/NECA/RRL

markup languages are designed to allow non-experts create multimodal presentations easily, representation languages in the above definition are designed to ease the integration and exchange of components in multimodal systems. Rist (2004) argues though that with the advent of more and more sophisticated authoring tools in the future, simplicity eventually will become less of a design criterion for markup languages and the distinction between the different types of representation languages will become more vague.

3.2 Scope – General Purpose Versus Application-Specific

To give an assessment of different strands of endeavours, we present, in the following, attempts to develop representation formats of broad scope and contrast them with languages that have been developed to serve much more restricted purposes, either being developed in the contexts of and thus particularly geared towards certain ECA implementations or aiming at the representation of certain aspects in multimodal behaviour generation.

3.2.1 General-Purpose Initiatives

HumanML (Brooks, 2000) was an initiative hosted by the "Organization for the Advancement of Structured Information Standards" (OASIS) to come up with a mark-up language for describing virtually all properties not only of artificial characters but of human beings. It set off with the goal to provide information for human-to-human and human-to-machine communications in a machine-readable form. The language aimed to encode information related to human communicative behaviour from high level (culture, emotion) to low level (signal, kinesics) and aimed to be of relevance for such diverse areas as anthropology, medicine, business communication and virtual reality. It was planned to specify tags related to physiology, proxemics, kinesics, haptics, beliefs, intentions, emotions, etc. and aimed to provide attributes related to community, culture, context/location of the conversation, personality, thoughts and signals. The initiative came to a halt soon after proposing a rough XML scheme with place-holders for the high-level concepts which never were specified in more detail.

Virtual Human Markup Language (VHML) presents a more down-to-earth initiative for a language that should facilitate the interaction of a talking head or a virtual human (Beard and Reid, 2002). It was designed as a confederation of various relatively simple sub-languages, each of them concerned with a sub-task: dialogue management, emotion, facial animation, body animation, hypertext and speech. VHML has a hierarchical structure, i.e. elements of a lower level inherit information from the higher level. The one typical example for the hierarchical encoding is emotion tags which are inherited by all the sub-components for speech, face and gesture. The specification of VHML did not leave the draft level and was mainly geared towards the control of a talking head, for which also sample implementations were implemented. That is, the most detailed specification was available for the head and the face, while the gesture markup language only comprised a small set of six atomic behaviours (shrug, agree, disagree, concentrate, emphasise, smile).

3.2.2 Special-Purpose Applications

Languages that have been developed for specific purposes are, e.g., SiGML for sign language (Elliott et al., 2004) and MURML for the reproduction of gesture kinematics (Kranstedt and Kopp, 2002). Other examples are MPML for a presentation agent (Mori et al., 2003); RRL to represent information relevant at the interfaces of system components in a pipeline for generation of animated presentation dialogues (Piwek et al., 2002); APML (De Carolis et al., 2004), AML and CML (Arafa et al., 2002) for agent communicative behaviour; and BEAT for verbal and non-verbal synchronisation (Cassell et al., 2001). We will describe MPML, RRL and APML in greater detail directly in the following. MURML will be addressed when it comes to gesture coding (Sect. 4.3.2) and BEAT when we talk about multimodal natural language generation (Sect. 4.2.1).

Multimodal Presentation Markup Language (MPML) aims at developing a language to easily create animated agents within interactive presentations. Agents may be set up on the web and the user can interact directly with the agents. The general goal of MPML is that, unlike most other web agents for presentation applications, the presentation of information is no longer presented sequentially, but its content is generated dynamically as the conversation between the agent and the user evolves (Mori et al., 2003). Furthermore, MPML has been designed for mouse control, voice control, text-to-speech and agent's action description (Tsutsui et al., 2000). A specialised scripting language, SCRipting Emotion-based Agent Minds (SCREAM), may be interfaced with MPML (Prendinger et al., 2004). SCREAM has been designed to create emotionally and socially appropriate responses of animated agents placed in an interactive environment. SCREAM is specialised in scripting the agent's mind. SCREAM may be used within applications that compute the verbal content of the interaction between the user and the agent. The role of SCREAM is to compute the emotion that may arise during the conversation. The instantiation of the signals and their intensity for a given emotion is then computed taking into account many factors, such as the social setting of the conversation as well as the agent's mental state.

The Rich Representation Language, RRL, has been developed to manage interactive dialogue scenes between two or more virtual agents (Piwek et al., 2002). RRL is used as a link between a scene generator, a multimodal natural language generator, a speech synthesis component, a gesture assignment component and finally a media player. A scene description contains information related to the set of communicative acts and the temporal ordering of these acts. An affective reasoner is embedded in the scene description to compute the corresponding emotion that may be triggered by given acts. The emotion is defined by its type, intensity and optionally by the object that causes the emotion. A scene description is input to the multimodal natural language generator that computes the corresponding linguistic and non-linguistic forms of the communicative acts. The role of the speech synthesis and gesture assignment components is to instantiate the acoustic and visual data for a given emotion and dialogue act. This results in an XML script representing the multimodal dialogue where the verbal and non-verbal behaviour is fully specified and the integrated temporal specification of the various communication channels (speech, facial, expression, gesture) is absolute. The script forms the common basis from which the specific animation directives required to drive individual players are derived via syntactic transformation.

The Affective Presentation Markup Language (APML) is based on a taxonomy of communicative functions proposed by Isabella Poggi (Poggi et al., 2000). A communicative function is defined as a pair (meaning, signal) where the meaning corresponds to the communicative value the agent wants to communicate and the signal is the behaviour used to convey this meaning. Communicative values are differentiated into four categories namely information about the speaker's beliefs, intentions, affective state and mental state. APML tags correspond to the meaning of a given communicative function. The conversion from meaning to signals is done by looking up a library of meaning–signal pairs.

Noot et al. (2004) developed GESTYLE, a complex representation language which is based on several dictionaries. Each dictionary reflects a certain aspect of the style of a character, e.g. cultural characteristics, profession or personality, and defines the association of meaning to signals. In addition physical information such as manner of gesturing (smooth, jerky, etc.) or tiredness can be specified. To create an agent with style one needs to specify this set of parameters (e.g. an Italian extrovert professor), and the proper set of mappings between meanings and signals is then instantiated.

To summarise, as we can see from the examples in this section, all proposed languages somehow model the relation between the ECA's mental states (goals, beliefs, intentions, emotions) and their display via concrete verbal and non-verbal behaviours. The details addressed in the various languages, however, differ widely, as they depend on the particular application the ECA system is built for, and thus on the system components realised.

4 Inventory of Information Relevant in the Generation Process

As we have seen in the previous section, all the languages or representation formats presented so far model some specific mix of information required in the complex process of multimodal behaviour generation. Obviously there is a considerable overlap between different representations, but at the same time they are not directly compatible or cannot be easily adapted for the needs of individual ECA projects. Therefore in the past we have seen (re-)building of representations over and over again. In this section, we will take a step back and concentrate not so much on existing languages, but look at the kinds of information relevant at different steps of the generation process.

Figure 1 depicts one way to conceptually organise the multitude of different processing steps and associated modules into fairly high-level blocks and helps to

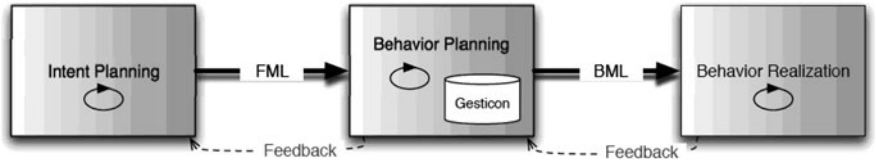

Fig. 1 Outlay of the processing pipeline for the generation of ECA behaviour as proposed by the SAIBA framework (Kopp et al., 2006)

divide the overall task into separate sub-components. This grouping into three top-level modules namely intent planning, behaviour planning and behaviour realisation has been proposed by the SAIBA framework Situation, Agent, Intention, Behaviour and Animation (Kopp et al., 2006; Vilhjálmsson et al., 2007). All three boxes have to be understood as complex systems with a variety of sub-components. One of the main guiding principles in the design of SAIBA architecture was to aim for a clear distinction between function and behaviour. FML thus stands for Functional Markup Language and BML for Behaviour Markup Language. Functional markup is to include all information regarding agent's mental, communicative and affective state that is necessary to create a link between intent and behaviour planning. It needs to provide a large spectrum of information including semantic, communicative, discursive, pragmatic and epistemic information. Behaviour markup comprises all those representations that are necessary for the realisation of behaviour. This includes textual and prosodic information, facial display, gestures and postures, eye gaze, etc. and, very importantly, it includes directives for the temporal synchronisation of behaviours. This format nevertheless aims for specifying behaviours independent of specific behaviour realisation systems, most specifically independent of concrete animation engines. In the following we are organising the discussion on information requirements and existing representation format along the lines of the SAIBA architecture.

4.1 Intent Planning

In order to be able to specify an ECA's communicative behaviours, first of all the underlying intent needs to be determined. At this stage, the basic semantic units related to the communicative event are computed. There is no reference to any physical or verbal behaviour yet. All in all, intent planning requires the computation of the mental, affective and communicative state of the agent. One possibility to implement the mental state of the agent is a BDI approach (Belief, Desire, Intention; Louis and Martinez, 2007). FML has not yet been defined in such a detail as BML has been. However, several proposals have been made for relevant concepts to be modelled during intent planning; see the collection of papers in Heylen et al. (2008). Contributions cover amongst others the specification of communicative actions (Kopp and Pfeiffer-Leßmann, 2008), different cognitive functions such as remembering or recalling (Mancini and Pelachaud, 2008), planning and regulating conversation (Lee et al., 2008), as well as emotional states (Krenn and Sieber, 2008).

4.1.1 Personality and Emotion

Personality and emotion are important aspects guiding the display of human behaviour. Emotions and emotion related states influence the way communication proceeds, its wording, voice quality, facial expression and other aspects of bodily behaviours such as posture and the dynamics of gesture. While emotion is responsible for the temporary changes in the quality of expression, personality determines the global tendencies of an individual's expression and thus may function as a means to establish coherency and consistency in the behaviour of an individual so that it becomes more predictable for the observer (Ortony, 2003). The expression of joy, for instance, in an extrovert person is overall much more pronounced than when expressed by an introvert person.

In the ECA community two approaches to personality and emotion are widely employed, the Five-Factor (OCEAN) model (Wiggins, 1996) and the OCC model (Ortony et al., 2003), respectively. This is reflected in the inventory of several representation languages for ECAs, e.g. PAR (Allbeck and Badler, 2002) makes provisions for OCC and OCEAN; MPML incorporates labels for the 22 emotions defined in OCC (Zong et al., 2000); RRL encodes the OCC labels plus their extension by (Elliott, 1992), a subset of OCEAN labels and a politeness attribute; APML defines its own set of emotion labels geared towards the communication situation (medical counselling) and the type of ECA (a talking head) used.

OCC is an example of an appraisal model. It defines emotions as positive and negative reactions to events, to other characters' or people's actions and to objects. Events are evaluated with respect to their desirability, actions according to their praiseworthiness, and objects in terms of their appeal to the agent. The subjective appraisal of a situation depends on the goals, standards and attitudes of the agent. Whereas attitudes are long-term affective states, emotions have a strong onset, but diminish over time. The latter is typically modelled via a decay function (Gebhard et al., 2003). Gebhard (2005) has presented a computational model that integrates emotion, mood and personality, standing for short-, mid- and long-term aspects of affect, respectively. As the terminology to describe human emotional life is manifold, we would like to redirect you to the opening chapter of the handbook (Cowie et al., 2010) for a discussion of concepts and definitions.

Modelling of emotion comes into play in intent and behaviour planning as well as in behaviour realisation. Whereas in intent planning, appraisal models have shown to be well suited (Ortony et al., 2003; Egges et al., 2003; Gebhard et al., 2003; Gebhard, 2005; Rank and Petta, 2005; Ochs et al., 2008; Marsella and Gratch, 2009), basic emotion categories (Ekman, 1993) are still the predominant representation when it comes to behaviour realisation. As Ekman's original research on basic emotions has focused on facial expressions, it unsurprisingly is still very influential in the field of facial animation for ECAs. Alternatively dimensional models

(Scherer, 2000) have been successfully employed for modelling emotional speech (Schröder et al., 2001; Schröder, 2004).

Personality models have been integrated in agents to model behaviour tendencies as well as intent planning (Moffatt, 1995; André et al., 2000; Johns and Silverman, 2001; Ball and Breese, 2000; Egges et al., 2004; Kshirsagar and Magnenat-Thalmann, 2002). The Five-Factor model of personality (McCrae and Costa, 1996) is used in most of the cited works. The interplay between personality and emotion has been studied. Moffatt (1995) views personality and emotion as similar states that differ in time and duration. Moreover personality ensures coherency of reactions to similar events, i.e. the emotional answers of an individual to these events are coherent through time. See also Ortony (2003).

4.1.2 Dialogue

Modelling of affect and personality not only is required for generating believable non-verbal behaviour, but also influences the agents dialogue; see Piwek (2003) for a survey. Even more importantly, automatic dialogue generation requires a representation of the domain, as it determines to a large extent what the virtual actors can talk about. Moreover behaviour, including dialogues, adheres to social conventions. Depending on the social relationship between the communication partners and the formality or informality of the situation, things are said differently and different display rules for the body behaviour apply (Walker et al., 1996; Rehm and André, 2005; De Carolis et al., 2001; Niewiadomski and Pelachaud, 2007; Ball and Breese, 2000), e.g. people would normally avoid crying in a business meeting whereas such a behaviour is much more likely in a private, intimate setting. What kinds of behaviours are socially acceptable and which ones are not strongly depends on cultural conventions. Just think about what is considered as good table manners in Europe as opposed to China. For instance, smacking and burping while eating will be considered as rude in Europe, but may be expected in China as an indicator for the positive appreciation of the food.

For modelling the communicative state, some dialogue planner is required that generates the initial version of a dialogue as a sequence of dialogue acts. A dialogue act represents an abstract communicative function, such as requesting for information, answering a question and giving feedback. Such communicative functions can be realised in many different ways depending, for example, on the personality and affective state of the actor. The structures produced by the dialogue planner represent communicative strategies that can be observed in a particular genre or domain and (partial) plans of how the communication should proceed. These plans include choice points according to which the communication differently proceeds based on the input from the outside world. This can be utterances from the communication partner(s), but also events occurring in the environment. van Deemter et al. (2008), e.g., describe the plan generation process for whole scenes of car sales dialogues enacted by two virtual characters, a customer and a seller.

This kind of presenter agents (see also André et al., 2000) realises very specific cases of dialogue where the whole dialogue is planned in one go depending on the initial settings given by the user. While the dialogue proceeds, no interference from the user is possible. Carmen's bright ideas (Marsella, 2003), FearNot! (Hall et al., 2006) and Faade (Mateas and Stern, 2003), are examples where the way how the story proceeds depends on the user's contribution to the dialogue.

4.2 Behaviour Planning

Given a particular intention and/or emotional state the agent aims to communicate, the system needs to decide which non-verbal signals will be used. The behaviour planner takes as input a given intention and/or emotional state, for instance to greet the communication partner happily, and outputs representations for the visual and acoustic signals to be generated by the behaviour realisation modules, such as a waving hand gesture, a broad smiling face and a greeting utterance for instance "Hello my friend! It is great to see you!". The behaviour planner has to instantiate the communicative acts, in our case greeting with the emotional colouring happy, which were generated during the intent planning phase and which are defined in terms of their meaning, into signals and then decide which modalities (facial expression, gaze, gesture, posture, voice, etc.) will be used to convey the particular meaning. Apart from selecting the modalities to convey meaning, proper synchronisation between the modalities is crucial. Examples of ECA systems that use such an approach are, for instance, the Greta behaviour engine (Pelachaud, 2005), BEAT (Cassell et al., 2001), SmartBody (Thiébaux et al., 2008) and MAX (Kopp and Wachsmuth, 2004). The task of behaviour planning thus at least comprises the sub-tasks multimodal planning and multimodal alignment.

If you go back to our example 2 on page 3, multimodal planning accounts for modelling the communication channels face, gaze, gesture and utterance expressed in c1 to c5, whereas multimodal alignment takes care of the timing of the expressions in the different channels relative to each other, which we have modelled with the start–end mechanism. Absolute timing is only available at the stage of behaviour realisation, when the speech is synthesised and concrete animations have been selected.

4.2.1 Multimodal Behaviour Planning

Non-verbal and verbal behaviour needs to be tightly integrated. Thus, in the most sophisticated ECA implementations, planning of non-verbal behaviour is coupled with natural language generation (NLG) leading to multimodal natural language generation (MNLG). In NLG the overall generation task is traditionally divided into two separate phases. In strategic generation it is decided "What to say", i.e. which propositions are to be expressed in a still language-independent representation. Tactical generation then deals with the "How to say", i.e. it is responsible to come up with the concrete wording.

As in NLG, the multimodal generation process is divided into a planning phase where the communicative acts are semantically outlined and a realisation phase where the behaviours that are going to be actually displayed are specified. Gestures are planned on the basis of the semantic and pragmatic content of the natural language utterances and are aligned with the respective representations of the utterance. At this stage, information on the concrete realisation of the body behaviour as well as the surface realisation of the utterances, i.e. the concrete wording, is still under-specified. The idea of intertwining gestural and syntactic structure has been proposed in different works. Cassell et al. (2000) describe a mechanism for applying the SPUD natural language generator (Stone et al., 2000) to multimodal generation. SPUD makes use of the "Lexicalized Tree Adjoining Grammar" (LTAG) formalism (Joshi et al., 1975; Schabes, 1990) and integrates the natural language grammar with motion events. Integration of gestures and syntax is particularly suitable for gestures that can express semantic content and therefore present an alternative to linguistic expression of the same content. For instance, if one wants to express that some X has a square shape one could say "X is squared", "X is of square shape", etc. However, one could as well say "X looks like this" and produce a gesture depicting a square. Gestures can also be used to express discourse functions. For instance, a question can be accompanied by an eye brow raise or a head tilt, assertions by a head nod, etc. van Deemter et al. (2008) describe an approach to MNLG which is based on typed feature structures representing deep syntactic, semantic and pragmatic content of dialogue acts and referring expressions. For the semantic representations, Discourse Representation Theory is employed (Kamp and Reyle, 1993). An extra module associates gestures and body postures with specific dialogue acts.

4.2.2 Alignment of Multimodal Behaviours

At the stage of behaviour planning signals across modalities must be aligned to each other. The most prevalent task is the temporal synchronisation between verbal and non-verbal signals, i.e. typically between eye gaze, facial expression, gestures and speech. Emphasis, e.g., is usually encoded via the synchronisation of an eye brow movement and a beat gesture with the stressed syllable of the word to be emphasised. Also speech, gaze and deictic gestures are typically synchronised. Take for example an utterance such as "give me this cake", which might be accompanied by an eye gaze and a deictic gesture towards the cake, with gaze and pointing being aligned with the phrase "this cake". This example shows that apart from the temporal alignment across channels, also the spatial alignment of movements and objects in the world needs to be handled. Getting back to the example, there is the cake as a target of pointing and gaze, and there is the addressee of the utterance who also needs to be looked at in order to establish and maintain the communication channel. In order to enable the behaviour realisation modules to take care of the actual synchronisations, the representations on the level of behaviour planning should make provisions for appropriate synchronisation points. Knowledge about the location of stressed syllables and sentence accents, e.g., is most crucial for proper synchronisation in the speech modality. For gestures usually the exact placement of the stroke phase, which is defined as the most meaningful part, is essential. In Fig. 2 the especially fine-grained inventory of synchronisation points offered by BML is depicted. The purpose of the different points is explained as follows.

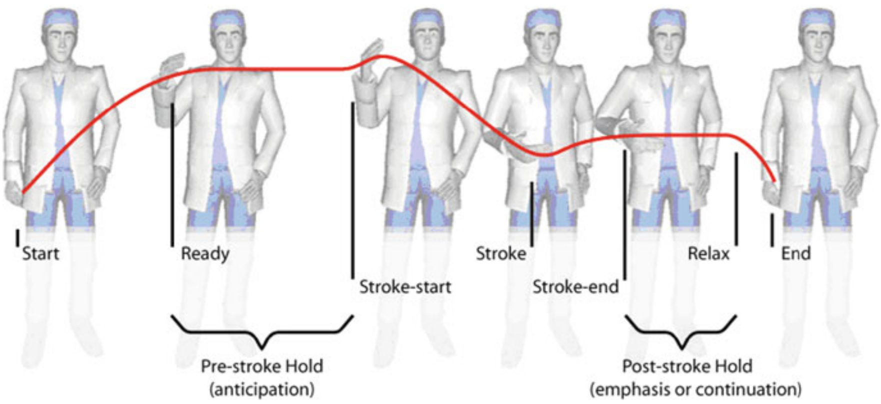

Fig. 2 The synchronisation points of a communicative behaviour (Kopp et al., 2006)

> The preparation for or visible anticipation of the behaviour occurs between start and ready, and the retraction back to neutral or previous state occurs between relax and end. The actual behaviour takes place between the ready and relax, with the most significant or semantically rich motion during the stroke phase, between stroke-start and stroke-end, with the greatest effort coinciding with the stroke point. (. . .) If no preparation or relaxation is needed, then start and ready refer to the same point in time, and relax and end refer to the same point in time. Quoted from Kopp et al. (2006)

4.3 Behaviour Realisation

Behaviour realisation is concerned with the generation of the concrete realisation of the behaviours. This component deals with tasks such as selecting the one most appropriate deictic gestures from a repository of candidate gestures that shall be finally realised by the media player, realising a very specific facial expression, generating speech on the basis of natural language text and replacing the specification of relative timing of the synchronised behaviours from the planning phase with absolute time values. Until today, it is common in ECA implementations to use speech as the guiding medium for temporal alignment. In the best case the granularity of temporal information in the speech channel goes down to the level of phoneme durations (in milliseconds), though in many applications only information on the location of word boundaries is employed. With the availability of fine-grained prosodic information, facial and gesture animation can be time-aligned to individual phonemes, accented syllables or boundaries of intonation phrases.

What is still missing is an integrated approach where not only speech defines the timing of its accompanying facial expressions and gestures, but also motor activation constrains voice quality, e.g. to lengthen the duration of a prosodic phrase and postpone the onset of its successor in order to wait for an accompanying gesture to finish.

4.3.1 Speech

Text-to-speech (TTS) systems take as input text possibly annotated with additional information. The TTS determines pronunciation and prosodic properties, such as location and type of pitch accents, prosodic phrase boundaries and duration of phonemes. Based on this information sound files are generated. For generating multimodal behaviour the TTS should provide fine-grained temporal information such as the list of phonemes, the location of phoneme boundaries to allow for the synchronisation of visemes (mouth shapes related to sounds/phonemes) or the location of accented syllables to allow for the exact temporal alignment of beat gestures. Though this information at some point is available within virtually every TTS system, there is no standard way to gain access to this information, given it is accessible at all. Unfortunately many of the commercial products do not provide interfaces to this sort of information, and only a few of the research-related systems provide access as easily as, e.g., the Mary system (Schröder and Trouvain, 2004) or the Festival TTS (Black and Taylor, 1997).

The W3C Speech Synthesis Markup Language SSML[4] has been developed to assist the generation of synthetic speech. It provides a standard way to mark up text in order to control aspects of speech such as pronunciation, volume, pitch and rate and is supported by a variety of speech synthesis systems. SSML can thus be seen as an example of a success story when it comes to the specification of a unimodal markup language. This standard, however, only provides a very rough interface both to and from the synthesis engine. For instance the only feedback mechanism is the insertion of event throwing tags in the text, which limits the temporal granularity of the feedback to the level of orthographic words. For the purpose of multimodal generation, in addition to SSML, some standardised speech synthesis output format would be highly desirable in order to make TTS's internal decisions on pronunciation, timing, pitch, accenting, phrasing, etc. accessible to other components.

As previously mentioned, speech is usually the leading modality providing the timeline for the other modalities (face, gaze, body, gesture) to align with. This, however, requires to be changed in favour of models where the timing of speech should be sensitive to restrictions from other modalities posing additional challenges for the specification of proper interfaces from and to the speech synthesis component. One example is the generation of explanations while manipulating objects, where the manipulations become the leading modality and pauses between intonation phrases are adapted accordingly. In this case, control of sub-sentential chunks has to be guaranteed (Kopp and Jung, 2000). Another possible application where a demand of increased temporal control is to be foreseen is the implementation of immediate reactivity, e.g. an ECA that is interrupting an utterance as an instant reaction to a barge-in or to some other observed user behaviour.

[4] http://www.w3.org/TR/speech-synthesis/

4.3.2 Gestures

Gestures are complex, being composed by one or a sequence of basic gesture elements, each of which describe a basic hand–arm movement trajectory. A trajectory is defined by a sequence of key points where each point corresponds to a position of the wrist in 3D space and a hand configuration. Depending on the granularity of representation, a gesture spans one or more phases, e.g. preparation, stroke, hold, retract. See also Fig. 2. Methods for encoding of gestures can be classified on a continuum ranging from purely semantic representations (related only to the meaning of the gesture) to formats which encode the form of the gesture exclusively. Most existing representation languages for computational systems are founded upon annotation systems in psychology and sign language research. In the following, we will briefly review some of the foundational works and then give an outline of scripting languages used for ECA systems. A more detailed review of existing gesture coding schemes can be found in Serenari et al. (2002). McNeill (1992) provides a semantic classification into iconic, metaphoric, deictic and beat gestures. To localise gestures, a grid-like gesture space is introduced in front of the actor. The descriptions of gesture form are holistic-imagistic though and not readily adaptable to automatic processing, because the shape of a gesture, especially in the case of iconic and some metaphoric gestures, refers to the meaning of the gesture. For example, a cup-shaped hand in certain contexts (for instance when visitors come to the house in France) carries the meaning "to offer something to the guests" as it is interpreted as representing a bowl of food. To make use of this representation for behaviour realisation, it requires the instantiation of the rough semantic classification into concrete representations, such as shape descriptions. Calbris (1990), thus, describes gestures by the meaningful form of their components morphology (segment, configuration, orientation, localisation and movement). Components may be linked to physical properties, for example, the flat vertical hand held between the speaker and the listener to mean stop symbolises the erection of a wall between the speaker and the listener as to show refusal. This act of refusal can also be done by throwing the head backward or even turning the head away. Calbris (1990) describes how gesture variants can be gathered as a class of gestures carrying one meaning. In contrast a gesture may be associated to several meanings depending on the contexts it occurs in. Stokoe (1978) introduces the concept of breaking gestural configurations down into formational parameters such as location and orientation of the wrist and hand shape. A hand shape is described by a thumb orientation and shapes of the other four fingers. The more recent HamNoSys notation framework (Prillwitz et al., 1989; Hanke, 2004), originally developed for sign languages, follows this breakdown into formational parameters and provides a dictionary of the most frequently used configurations. MURML (Kranstedt and Kopp, 2002) is an XML-based description language which has been influenced by HamNoSys. It allows for detailed control of parallel and sequential components of gestures, whereas hand shapes and facial expressions are specified by simple labels. Several parameters have been defined such as the wrist location in space, the palm and finger orientation, hand shape and the wrist trajectory. Values of these parameters create hand and arm configuration.

Such configurations must be described for each phase of a gesture (preparation, pre-stroke hold, stroke, etc). Timing information related to the duration of a gesture to the temporal constraints between body parts involved within the gesture can be provided. MURML allows for a precise description of behaviour. Hartmann et al. (2002) describe a language for ECA animation that unites features of McNeill and HamNoSys. FORM2 by Martell et al. (2003) is an annotation scheme that captures the exact configuration and orientation of the arms and hands of a gesturer. It focuses on a detailed description of the geometrical properties of gestures. A corpus of annotated video material is available.

Scripting Technology for Embodied Persona, STEP (Huang et al., 2004), works for ECAs based on H-Anim.[5] It offers a set of sensors and effectors through which agents can perceive the world they are placed in and can take appropriate actions. STEP includes two main primitive actions (turn, move) to specify body movement. The first defines the rotation to apply to a given joint of the virtual agent, while the second relates to the displacement of one effector. More complex actions can be obtained by combining these primitive actions with three operators: seq, the sequential operator; par, the parallel operator; and T, the repeated operator.

4.3.3 Facial Expression

As with the other kinds of information relevant in multimodal behaviour generation, several coding schemes to describe facial expressions have been devised too. MPEG-4, for instance, is an ISO/IEC standard which defines specifications for the animation of face and body models within an MPEG-4 terminal (Doenges et al., 1997; Ostermann, 1998; Pandzic and Forchheimer, 2002). Two sets of parameters describe and animate the 3D facial model: the facial animation parameters (FAPs) and the facial definition parameters (FDPs). FDPs define the shape of the model, FAPs define the facial actions. FAPs represent a large set of basic facial actions including head, tongue, eye and mouth movement. In combination they represent facial expressions. Facial expression may also be coded using the Facial Action Coding System (FACS) developed by Ekman et al. (Ekman and Friesen, 1978; Ekman et al., 2002). It is a framework to measure facial signals using minimal action units (AUs). With FACS facial action units can be encoded on a scale of five intensities. Behaviour changes along this scale are carefully described. Paradiso and L'Abbate (2001) have established an algebra to create facial expressions. The authors have elaborated operators that combine and manipulate facial expressions. Another definition language has been proposed by De Carolis et al. (2004) and by Paradiso and L'Abbate (2001). In their language, an expression may be defined at a high level (a facial expression is a combination of other facial expressions already pre-defined) or at a low level (a facial expression is a combination of facial parameters). The low-level facial parameters correspond to the MPEG-4 FAPs. The language is also suitable to create easily extendable facial display dictionaries.

[5] http://www.h-anim.org/

5 Towards a Common Framework for Representations in Multimodal Behaviour Generation

The non-exhaustive overview above provides an impression on the amount of effort that has up to now been invested in the design of representation formats. In the last decade we have seen the development of a significant number of ECA systems and applications. In parallel, we have experienced the publication of an almost equal number of usually XML-based markup and representation languages – cf. Arafa et al. (2003) for an impressive list of such languages. To some extent the proliferation of representation languages can also be explained with the enthusiasm about XML in its early hey-days. Employing XML came with an implicit promise of reusability and ease of application and, in addition to presenting ECA systems, the associated representation formats suddenly became a topic of interest worth publishing. Though there are identifiable similarities between existing representation languages still there are not many examples where a sharing of representation formats – not to speak of software – has taken place. Considering the work that has been put to this topic, there is a significant lack on formats that ever got reused outside their original institution or ECA project. This lack of success in terms of acceptance in the community is most obvious for VHML and especially HumanML, which were explicitly designed with the aim to become standards. For other languages reusability might not have been the primary goal in the first place, as they were designed to fit the needs of a concrete ECA application. But even if the main motivation for existing representations might not have been to trigger and sustain the development of reusable and exchangeable system components via the specification of open interface formats, there is an obvious demand for such formats. The implementation of ECA systems requires expertise in such diverse research topics as emotion modelling, behaviour planning, natural language generation, speech synthesis and computer animation. Only for some of these tasks off-the-shelf modules are available; other components are still in their early stages of research and development. Given the overall complexity of ECA systems, exchangeable sub-components that would allow for a plug-and-play approach for the system development would clearly be desirable. Due to the lack of common standards and architectures, similar functionalities need to be implemented over and over again for different systems. At a second thought the problem can also be stated the other way round: It is not so much the lack of standardised interfaces which prevents the development of common software modules, but the lack of widely used system components hinders the establishment of common representation languages. There is a mutual dependency between interface specifications and the availability of system components, but of course software modules that are both useful and usable for a broader group of developers do indeed provide a high incentive to promote the interface formats connected with these components. These considerations have led to a renewed interest in the development of common interfaces, which should then foster the development of system components that could be shared and reused among different ECA projects, thus avoiding the replication of effort. An exemplary activity in this direction is the work of the SAIBA initiative where several research groups have joined forces to come up with a commonly agreed on framework for multimodal behaviour generation. The goal for this initiative is to develop common specifications for representation languages, which are meant to be application independent and graphics model independent, and to present a clear-cut separation between information types (function versus behaviour specification). Intermediate results of this joint endeavour are documented in Kopp et al. (2006) and Vilhjálmsson et al. (2007). Experience with earlier initiatives in this direction like VHML provides strong evidence that success or failure of such a representation format is tightly coupled to the availability of software components that actually provide an immediate benefit for the system developer. This of course is a chicken or egg problem not easily solved. We thus are re-evaluating exemplary sub-topics for which we think that shared representations and jointly developed software modules could succeed in the intermediate future. It is not by chance that the work within the SAIBA initiative has by now mostly focused on issues concerning the specification of BML, i.e. on schemes that deal with the encoding of behaviours. Though still complicated enough, there is a joint understanding of the concepts necessary to describe the communicative actions of the human body. Human anatomy and the specifics and needs of existing animation techniques and speech synthesisers help to guide the development of BML. For FML, i.e. the encoding of functional categories, it is much more difficult to come up with general, application-independent representations. Among the information types affiliated to the functional domain in SAIBA, the representation of emotions is a prime candidate for the development of a joint representation format; see the chapter "Representing Emotions and Related States in Technological Systems" by Schröder et al. (2010) in this part.

One important issue in the current development of BML is the specification of non-verbal communicative behaviour. Existing schemes which are using joint-angles and segment translations such as MPEG-4 or BVH (Pandzic and Forchheimer, 2002) do provide exact and detailed physical information on body shapes. Nevertheless they are viewed as way too specific. They are lacking flexibility both in the modification of temporal and spatial properties, e.g. they run into problems of collision if body proportions are changed. Also functional information, e.g. the identification of stroke phases, which is crucial for the proper temporal alignment across modalities, is completely missing. Coming up with a common higher- level format for the representation of facial expression, gesture and posture which would function as a sort of middle layer between the specification of intentions and the formats used for actually rendering body movements is by no means a trivial task. But still the problem seems to be confined in a tractable way. There is much agreement on the overall requirements for such a representation, and the development can partly be based on experience gained with existing languages like MURML and FORM. The prospect of coming up with a representation for the physical appearance and bodily actions of an ECA that is independent of individual animation engines is obviously appealing. A strong motivation for working on such a representation for behaviours has been the idea to use it for providing repositories of communicative gestures, for which the terms gesticon and gestuary have been proposed (Krenn and Pirker, 2004), i.e. collections of gestures reusable in different

ECA systems. At the same time the immediate requirement for supplementing the representation format with concrete software that actually provides non-trivial functionalities becomes evident. In the terminology of the SAIBA architecture this would be, e.g., behaviour realisation modules which are interpreting BML representations and render them to different animation engines. As long as no components for interpreting and transforming to player-specific code are available, this intermediate representation does not provide any additional functionality and developers of ECA systems would skip it and stick to their own representations. When it comes to representation formats for speech synthesis, we are facing a special situation. Speech synthesis is the one domain where exchangeable off-the-shelf components actually do exist. Basically all these systems provide an input interface and an output interface that are universally accepted, namely text and audio files, respectively. Even if developers of ECA systems might not always be happy with the quality of the synthesiser's output, these systems deal with a clearly defined, specialised and complex task, and not many developers feel inclined to intermingle with this functionality themselves. Open issues one could think of when it comes to ECAs and speech synthesis are missing standards to specify emotion and the missing ability for incremental speech synthesis which would be desirable for the really interactive systems that would, e.g., react to interruptions by the user in mid-sentence. But there is another issue on representation formats and speech synthesis that is not so much a technical problem. In most implementations of multimodal systems, speech is the leading modality which provides the temporal grid to which the other modalities (lip movements, gestures) are synchronised. Information on the temporal locus of stressed syllables, phonemes, accents, etc. thus is crucial, but in most TTS systems this data is not made available to the user. This is not due to technical reasons, but this information is usually suppressed because of mere ignorance of an existing demand for it. The promotion of a standardised format for this kind of information, i.e. a kind of speech output format, could trigger an increased awareness of TTS suppliers for that kind of information demand by ECA developers. As stated above, not only does the fate of any representation format rely on design factors such as expressive power, flexibility and ease of application but its acceptance in a wider community also strongly depends on the availability of implementations that actually support the creation and interpretation of the proposed format. We are concluding this contribution with a short survey on possible insights which could either be gained from or shared with research domains outside the narrow ECA paradigm. Descriptive schemes used for ECAs of course have much in common with coding schemes that describe the behaviour, physical appearance or emotional states of real humans, and ECA developers have been adapting coding schemes originally designed for humans in the past. FACS (Ekman and Friesen, 1978; Ekman et al., 2002), the coding scheme for facial expressions, was developed in the context of psychological and anthropological research and was very influential for specifying codes for facial animations. HamNoSys (Hanke, 2004), a representation format for sign language, has been adapted for the encoding of gestures in ECAs. Descriptive schemes for gestures have been developed for the manual annotation of multimodal conversational data, e.g. ANVIL (Kipp, 2001) and CoGest (Trippel et al., 2004).

Demands for a common gesture description language have also been brought up in the field of automatic gesture recognition and interpretation (Kölsch and Martell, 2006). Another source for inspiration and possible synergies is the gaming industry. Yue and de Byl (2006) present a standardisation initiative for AI components used in games. They deal with problems that are related to the intent and behaviour planning and not so much to the behaviour realisation, e.g. path finding and steering in a game environment and action planning. On an abstract level these are functionalities that resemble those in the behaviour planning component in SAIBA. There are insights to be gained by the way this standardisation process is organised. One interesting aspect is that this initiative does not bother with XML formats but deals with the specification of Application Programming Interfaces (APIs), i.e. formulates their interfaces in directly implementation-related terms.

The gaming industry also provides examples on how pseudo-standards actually may emerge from the spreading of tools and vice versa. BVH (Biovision Hierarchy) is a graphics format developed for storing motion-captured data. Tools for creating key-frame animations in BVH are also emerging, e.g. Cal3D, an open source character animation library,[6] and plug-ins for exporting BHV format from widely used commercial graphics programs such as 3D Studio Max. The BVH format is also used for animating avatars in the virtual world of Second Life and is an example of how an application provides the incentive for the development of tools and thus for the proliferation of specific representation formats. It remains to be seen whether a similar momentum can be gained in the development of representation formats for ECAs in the future.

References

Allbeck J, Badler N (2002) Towards representing agent behaviours modified by personality and emotion, In: Marriott A, Pelachaud C, Rist T, Ruttkay Z, Vilhjalmsson H (eds) Embodied conversational agents: let's specify and compare Them!, workshop notes, autonomous agents and multiagent systems 2002, July 16, University of Bologna, Bologna, Italy

André E, Rist T (2000) Presenting through performing: on the use of life-like characters in knowledge-based presentation systems. In: Proceedings of the 2000 international conference on intelligent user interfaces, New Orleans, LA, USA, 9–12 January 2000

André E, Rist T, van Mulken S, Klesen M, Baldes S (2000) The automated design of believable dialogues for animated presentation teams. In: Cassell J, Sullivan J, Prevost S, Churchill E (eds) Embodied conversational agents. MIT Press, Cambridge, MA

André E, Klesen M, Gebhard P, Allen S, Rist T (2000) Integrating models of personality and emotions into life-like characters. In: Paiva A (ed) Affective interactions: towards a new generation of computer interfaces. Lecture Notes in Computer Science, Vol 1814, Springer, Berlin

Arafa Y, Kamyab K, Mamdani E, Kshirsagar S, Guye-Vuilléme A, Thalmann D (2002) Two approaches to scripting character animation. In: Marriott A, Pelachaud C, Rist T, Ruttkay Z, Vilhjálmsson H (eds) Embodied conversational agents: let's specify and compare them!, workshop notes, autonomous agents and multiagent systems 2002, July 16. University of Bologna, Bologna, Italy

[6] https://gna.org/projects/cal3d/

Arafa Y, Kamyab K , Mamdani E (2003) Character animation scripting languages: a comparison. In: Rosenschein JS et al. (eds) Proceedings of the second international joint conference on autonomous agents and multiagent systems (AAMAS 2003), July 14–18, Melbourne, Australia. ACM Press, New York, NY, pp 920–921

Ball G, Breese J (2000) Emotion and personality in a conversational agent. In: Cassell J, Sullivan J, Prevost S, Churchill E (eds) Embodied conversational agents. MIT Press, Cambridge, pp 189–219

Beard S, Reid D (2002) MetaFace and VHML: A first implementation of the virtual human markup language. In: Marriott A, Pelachaud C, Rist T, Ruttkay Z, Vilhjalmsson H (eds) Embodied conversational agents: let's specify and compare them!, workshop notes, autonomous agents and multiagent systems 2002, July 16. University of Bologna, Bologna, Italy

Bickmore T, Cassell J (2005) Social dialogue with embodied conversational agents. In: van Kuppevelt J, Dybkjaer L, Bernsen N (eds) Advances in natural, multimodal dialogue systems. Kluwer, New York, NY

Black AW, Taylor PA (1997) The festival speech synthesis system: system documentation. Technical Report HCRC/TR-83, Human Communication Research Centre, University of Edinburgh, Scotland, UK. http://www.cstr.ed.ac.uk/projects/festival/. Accessed 3 May 2010

Brooks R (ed) (2002) Human Markup Language Primary Base Specification 1.0, OASIS HumanMarkupTC. http://www.oasis-open.org/committees/download.php/60/HM.Primary-Base-Spec-1.0.html. Accessed 31 May 2010

Calbris G (1990) The semiotics of French gestures. University Press, Bloomington, IN

Cassell J, Pelachaud C, Badler N, Steedman M, Achorn B, Becket T, Douville B, Prevost S, Stone M (1994) Animated conversation: rule-based generation of facial expression, gesture and spoken intonation for multiple conversational agents. In: Proceedings of Siggraph 94, ACM SIGGRAPH, Addison Wesley, Massachu setts, pp 413–420

Cassell J, Bickmore T, Billinghurst M, Campbell L, Chang K, Vilhjálmsson H., Yan H (1999). Embodiment in conversational interfaces: rea. In: Proceedings of the CHI'99 Conference, Pittsburgh, PA, pp 520–527

Cassell J., Stone M, Yan H (2000) Coordination and context-dependence in the generation of embodied conversation. In: First international natural language generation conference (INLG'2000), June 12, Mitzpe Ramon, Israel, pp 171–178

Cassell J, Sullivan J, Prevost S, Churchill E (eds) Embodied conversational agents. MIT Press, Cambridge, MA

Cassell J, Vilhjálmsson H, Bickmore T (2001) BEAT: The behavior expression animation toolkit. In: Proceedings of SIGGRAPH '01, Los Angeles, CA, pp 477-486, August 12–17

Cowie R, Sussman N, Ben-Ze'ev A (2010) Emotions: concepts and definitions. In: this volume

De Carolis B, Pelachaud C, Poggi I, De Rosis F (2001) Behavior planning for a reflexive agent. In: Proceedings of IJCAI 2001, Oporto, Portugal, April, 2001

De Carolis B, Pelachaud C, Poggi I, Steedman M (2004) APML, a mark-up language for believable behavior generation. In: Prendinger H, Ishizuka M (eds) Life-like characters. tools, affective functions and applications, Springer, Berlin, pp 65–85

de Rosis F, Pelachaud C, Poggi I, Carofiglio V, De Carolis N (2003) From greta's mind to her face: Modeling the dynamics of affective states in a conversational embodied agent. Special Issue on "Applications of affective computing in human-computer interaction". Int J Human-Comput Stud 59(1–2): 81–118

Doenges P, Capin TK, Lavagetto F, Ostermann J, Pandzic IS, Petajan E (1997) MPEG-4: Audio/video and synthetic graphics/ audio for real-time, interactive media delivery, signal processing. Image Commun J. 9(4): 433–463

Egges A, Kshirsagar S, Magnenat-Thalmann N (2003) A model for personality and emotion simulation. In: Knowledge-based intelligent information and engineering systems. Lect Notes Comput Sci 2773/2003: 453–461

Egges A, Kshirsagar S, Magnenat-Thalmann N (2004) Generic personality and emotion simulation for conversational agents. J Visual Comput Animation 15(1): 1–13

Ekman P (1993) Facial expression of emotion. Am Psychol 48: 384–392

Ekman P, Friesen W (1978) Facial action coding system. Consulting Psychologists Press, Palo Alto, CA

Ekman P, Friesen W, Hager J (2002) Facial action coding system: the manual. A Human Face, Salt Lake City

Elliott CD (1992) The affective reasoner: a process model of emotions in a multi-agent system. Ph.D. Thesis, Northwestern University, Illinois

Elliott R, Glauert J R W, Jennings V, Kennaway J R (2004) An overview of the SiGML notation and SiGMLSigning software system. In: Streiter O, Vettori C (eds) 4th international conference on language resources and evaluation (LREC 2004), Lisbon, Portugal, May 26–28, 2004, pp 98–104

Gebhard P (2005) ALMA – a layered model of affect. In: Proceedings of the fourth international joint conference on autonomous agents and multiagent systems (AAMAS'05), Utrecht University, Utrecht July 25-29, 2005, pp 29–36

Gebhard P, Kipp M, Klesen M, Rist T (2003) Adding the emotional dimension to scripting character dialogues. In: Proceedings of the 4th international working conference on intelligent virtual agents (IVA'03), – Irsee, Germany, 15-17 September, 2003, pp 48–56

Hall L, Vala M, Hall M, Webster M, Woods S, Gordon A, Aylett R (2006) FearNot's appearance: reflecting children's expectations and perspectives. In: Gratch J, Young M, Aylett R, Ballin D, Olivier P (eds) 6th international conference on intelligent virtual agents (IVA'06), Springer, Berlin, LNAI 4133, pp 407–419

Hanke T (2004) HamNoSys; representing sign language data in language resources and language processing contexts. In: Proceedings of 4th international conference on language resources and evaluation (LREC 2004), Lisbon, Portugal, 26–28 May, 2004, pp 1–6

Hartmann B, Mancini M, Pelachaud C (2002) Formational parameters and adaptive prototype instantiation for MPEG-4 compliant gesture synthesis. In Proceedings of computer animation 2002 (CA 2002), Geneva, Switzerland, 19-21 June, 2002, p 111

Heylen D, Kopp S, Marsell S, Pelachsud C, Vilhjalmsson H (eds) (2008) Why conversational agents do what they do. Functional representations for generating conversational agent behavior, AAMAS 2008 Workshop 2, April 9, Estoril

Huang Z, Eliens A, Visser C (2003) XSTEP: a markup language for embodied agents. In: Proceedings of the 16th international conference on computer animation and social agents (CASA'2003), May 8–9, Rutgers University, New-Brunswick, NJ, USA, IEEE Computer Society, Washington, DC, pp 105–110

Huang Z, Eliens A, Visser C (2004) STEP: a scripting language for embodied agents. In: Prendinger H, Ishizuka M (eds) Life-like characters, tools, affective functions and applications, Springer, Berlin

Johns M, Silverman, BG (2001) How emotions and personality effect the utility of alternative decisions: a terrorist target selection case study. In: Tenth conference on computer generated forces and behavioral representation. SISO. Norfolk, Virginia, pp 55–64

Johnson JH, Vilhjálmsson H, Marsella S (2005) Serious games for language learning: how much game, how much AI? In: 12th international conference on artificial intelligence in education, Amsterdam, The Netherlands, 18–22 July, 2005

Joshi AK, Levy L, Takahashi M (1975) Tree adjunct grammars. J Comput Syst Sci 10: 136–163

Kamp H, Reyle U (1993) From discourse to logic. Kluwer, Dordrecht

Kendon A (1990) Conducting interaction. Cambridge University Press, Cambridge

Klesen M, Gebhard P (2004) Player markup language. Version 1.2.4, DFKI, internal document

Kölsch M, Martell C (2006) Toward a common human gesture description language, workshop on specification of mixed reality user interfaces: approaches, languages, standardization, IEEE Virtual Reality Conference (VR 06), Alexandria, VA, 25 March, 2006

Kipp, M (2001) Anvil – a Generic annotation tool for multimodal dialogue, In: Proceedings of the 7th European conference on speech communication and technology (Eurospeech), Aalborg, Denmark, 3–7 September (2001) pp 1367–1370

Kopp S, Jung B (2000) An anthropomorphic assistant for virtual assembly: max. In: Proceedings of the autonomous agents '00 workshop: communicative agents in intelligent environments, Barcelona, Spain

Kopp S, Wachsmuth I (2004) Synthesizing multimodal utterances for conversational agents. J Comput Animation Virtual Worlds 15(1): 39–52

Kopp S, Krenn B, Marsella S, Marshall A, Pelachaud C, Pirker H, Thórisson K, Vilhjálmsson H (2006) Towards a common framework for multimodal generation: the behaviour markup language. In: Gratch J et al (eds) Intelligent virtual agents 2006, LNAI 4133. Springer, Berlin, pp 205–217

Kopp S, Pfeiffer-Leßmann N (2008) Functions of speaking and acting. In: Heylen D, Kopp S, Marsell S, Pelachsud C, Vilhjalmsson H (eds) Why conversational agents do what they do. Functional representations for generating conversational agent behavior, AAMAS 2008 Workshop 2, April 9, Estoril

Kranstedt A, Kopp S, Wachsmuth I (2002) MURML: a multimodal utterance representation markup language for conversational agents. In: Marriott A, Pelachaud C, Rist T, Ruttkay Z, Vilhjalmsson H (eds) Embodied conversational agents: let's specify and compare them!, workshop notes, autonomous agents and multiagent systems 2002, University of Bologna, Bologna, Italy, 16 July, 2002

Krenn B, Pirker H (2004) Defining the gesticon: language and gesture coordination for interacting embodied agents. In: Proceedings of the AISB-2004 symposium on language, speech and gesture for expressive characters, University of Leeds, UK, March 29–April 1, 2004, pp 107–115

Krenn B, Sieber G (2008) Functional Mark-up for behaviour planning. Theory and practice. In: Heylen D, Kopp S, Marsell S, Pelachsud C, Vilhjalmsson H (eds) Why conversational agents do what they do. Functional representations for generating conversational agent behavior, AAMAS 2008 Workshop 2, April 9, Estoril

Krenn B, Grice M, Piwek P, Schröder M, Klesen M, Baumann S, Pirker H, van Deemter K, Gstrein E (2002) Generation of multi-modal dialogue for net environments. In: Proceedings of KONVENS-02, Saarbrcken, Germany, September 30–October 2 (2002)

Kshirsagar S, Magnenat-Thalmann N (2002) A multilayer personality model. In: Proceedings of 2nd international symposium on smart graphics, ACM Press, New York, NY, pp 107–115

Lee J, DeVault D, Marsella S, Traum D (2008) Thoughts on FML: behavior generation in the virtual human communication architecture. In: Heylen D, Kopp S, Marsell S, Pelachsud C, Vilhjalmsson H (eds) Why conversational agents do what they do. Functional representations for generating conversational agent behavior, AAMAS 2008 Workshop 2, 9 April 2008, Estoril

Louis V, Martinez T (2007) JADE semantics framework. In: Developing multi-agent systems with jade. Wiley, Chichester, pp 225–246

Mancini M, Pelachaud C (2008) The FML-APML language. In: Heylen D, Kopp S, Marsell S, Pelachsud C, Vilhjalmsson H (eds) Why conversational agents do what they do. Functional representations for generating conversational agent Behavior, AAMAS 2008 Workshop 2, 9 April 2008, Estoril

Marsella S (2003) Interactive pedagogical drama: Carmen's bright IDEAS assessed. In: Proceedings of the 4th international working conference on intelligent virtual agents (IVA'03), Irsee, Germany, 15-17 September, 2007, pp 1–4

Marsella, S. and Gratch, J (2009) EMA: A model of emotional dynamics. J Cogn Syst Res 10(1): 70–90

Marsella S, Johnson WL, LaBore C (2003) Interactive pedagogical drama for health interventions. In: Proceedings of the 11th international conference on artificial intelligence in education AIED 2003, Sidney, Australia, 20–24 September 2003

Martell C (2002) FORM: an extensible, kinematically-based gesture annotation scheme. In: Proceedings of ICSLP-2002, Denver, Colorado, 16–20 September, 2002, pp 353–356

Martell C, Howard P, Osborn C, Britt L, Myers K (2003) FORM2 kinematic gesture. Video recording and annotation. Linguistic Data Consortium LDC, Philadelphia, PA

Mateas M, Stern A (2003) Facade: an experiment in building a fully-realized interactive drama. In: Game Developer's Conference: Game Design Track, San Jose, CA, USA, 20–24, March 2003

Mateas M, Stern A (2004) A behaviour language: joint action and behavioural Idioms. In: Prendinger H, Ishizuka M (eds) Life-like characters. Tools, affective functions, and applications. Springer, Berlin, pp 19–38

Matheson C, Pelachaud C, de Rosis F, Rist T (2003) MagiCster: believable agents and dialogue. In: Künstliche Intelligenz, special issue on "Embodied Conversational Agents", November 2003, 4, pp 24–29

McCrae R R, Costa P T Jr. (1996) Toward a new generation of personality theories: theoretical contexts for the five-factor model. In: Wiggins SJ (ed) The five-factor model of personality: theoretical perspectives. Guilford, NY, pp 51–87

McNeill D (1992) Hand and mind – what gestures reveal about thought. The University of Chicago Press, Chicago, IL

Moffat D (1995) Personality parameters and programs. In: Lecture notes in artificial intelligence: creating personalities for synthetic actors: towards autonomous personality agents. LNCS. doi:10.1007/BFb0030565, pp 120–165

Moreno, R (2007) Animated software pedagogical agents: how do they help students construct knowledge from interactive multimedia games? In: Lowe R, Schnotz W (eds) Learning with animation. Cambridge University Press, New York, NY, pp 183–207

Mori K, Jatowt A, Ishizuka M. (2003) Enhancing conversational flexibility in multimodal interactions with embodied lifelike agents. In: Proceedings of the International conference on intelligent user interfaces (IUI 2003), Miami, Florida, 12–15 January 2003, pp 270–272

Niewiadomski R, Pelachaud C (2007) Model of facial expressions management for an embodied conversational agent. In: Proceedings of ACII 2007, September 12-14, Lisbon. LNCS. doi:10.1007/978-3-540-74889-2, pp 12–23

Nijholt A (2006) Towards the automatic generation of virtual presenter agents. In: Proceedings of InSITE 2006, June 25-28, Salford. Infor Sci 9: 97–115

Nischt M, Prendinger H, André E, Ishizuka M (2006) Creating three-dimensional animated characters: an experience report and recommendations of good practice. Upgrade: virtual environments 7(2): 35–41. http://www.upgrade-cepis.org/issues/2006/2/upgrade-vol-VII-2.pdf. (Accessed) 31 May 2010

Nischt M, Prendinger H, André E, Ishizuka (2006) MPML3D: a reactive framework for the multimodal presentation markup language. In: Proceedings of the 6th international conference on intelligent virtual agents (IVA'06), August 21-23, Marina del Rey, CA, LNCS. doi:10.1007/11821830, pp 218–229

Noot H, Ruttkay Z (2004) Gesture in style. In: Camurri A, Volpe G (eds) Gesture-based communication in human-computer interaction – GW 2003. LNCS vol 2915, Springer, Berlin, pp 471–472

Ochs M, Pelachaud C, Sadek D (2008) An empathic virtual dialog agent to improve human-machine interaction. In: Seventh international joint conference on autonomous agents and multi-agent systems, AAMAS'08, Estoril Portugal, 12–16, May 2008, pp 89–96

Ortony A (2003) On making believable emotional agents believable. In: Trappl R, Petta P, Payr S (eds) Emotions in humans and artefacts. MIT Press, Cambridge, MA pp 189–212

Ortony A, Clore GL, Collins A (1988) The cognitive structure of emotions. Cambridge University Press, Cambridge

Ostermann J (1998) Animation of synthetic faces in MPEG-4. In: Proceedings of the computer animation' 98, Philadelphia, PA, USA, 8–10 June 1998, pp 49–51

Pandzic IS, Forchheimer R (eds) (2002) MPEG4 facial animation – the standard, implementations and applications. Wiley, New York, NY

Paradiso A, L'Abbate M (2001) A model for the generation and combination of emotional expressions. In: Proceedings of the AA' 01 workshop on multimodal communication and context in embodied agents, Montreal, Canada, 29 May 2001

Pelachaud C (2005) Multimodal expressive embodied conversational agents. In: Proceedings of the 13th annual ACM international conference on multimedia. SESSION: brave new topics 2: affective multimodal human-computer interaction; Singapore, 6–11, November 2005, pp 683–689

Peltz J, Kumar Thunga R (2005) HumanML: The Vision. TheHumanMLReport-WhiteP, July 12. http://www.oasis-open.org/committees/download.php/13625/HumanMLReport-WhitePaper.pdf Accessed 31 May 2010

Piwek P (2003) An annotated bibliography of affective natural language generation. version 1.3. (version 1.0 appeared in 2002 as ITRI Technical Report ITRI-02-02, University of Brighton). http://www.itri.brighton.ac.uk/projects/neca/affect-bib.pdf Accessed 31 May 2010

Piwek P, Krenn B, Schröder M, Grice M, Baumann S, Pirker H (2002) RRL: a rich representation language for the description of agent behaviour in NECA. In: Marriott A, Pelachaud C, Rist T, Ruttkay Z, Vilhjalmsson H (eds) Embodied conversational agents: let's specify and compare them!, workshop notes, Autonomous Agents and Multiagent Systems 2002, University of Bologna, Bologna, Italy, 16 July, 2002

Poggi I, Pelachaud C, de Rosis F (2000) Eye communication in a conversational 3D synthetic agent. AI Commun 13(3): 169–182

Predinger H, Ishizuka M (eds) (2004) Life-like characters. Cognitive technologies. Springer, Berlin

Prendinger H, Saeyor S, Ishizuka M (2004) MPML and SCREAM: scripting the bodies and minds of life-like characters. In: Predinger H, Ishizuka M (eds) Life-like Characters. Cognitive technologies. Springer, Berlin, pp 213–242

Prillwitz S, Leven R, Zienert H, Hanke T, Henning J (1989) Hamburg notation system for sign languages: an introductory guide. In: International studies on sign language and communication of the deaf, vol 5. Signum Press, Hamburg, Germany

Rank S, Petta P (2005) Appraisal for a character-based story-world. In: Panayiotopoulos T et al (eds) Intelligent virtual agents, 5th international working conference, IVA 2005. Kos, Greece, September 12–14. Springer, Berlin pp 495–496

Rehm M, André E (2005) From chatterbots to natural interaction – face to face communication with embodied conversational agents. IEICE transactions on information and systems, special issue on life-like agents and communication. Oxford University Press Oxford, Oxford, UK, pp 2445–2452

Rehm M, André E (2005) Catch me if you can: exploring lying agents in social settings. In: Proceedings of the fourth international joint conference on autonomous agents and multiagent systems AAMAS '05. July 25 – 29, Utrecht, The Netherlands. ACM, New York, NY, pp 937–944

Rehm M, André E (2005) Informing the design of embodied conversational agents by analysing multimodal politeness behaviors. In: AISB symposium for conversational informatics, University of Hertfordshire, Hatfield, England, 12–15 April 2005

Rickel J, Johnson WL (1998) STEVE: a pedagogical agent for virtual reality. In: Sierra C et al (eds) Proceedings of the second international conference on autonomous agents (Agents'98), 9–13, Minneapolis/St. Paul, MN, USA. ACM Press, New York, NY, pp 332–333

Rist T (2004) Issues in the design of scripting and representation languages for life-like characters. In: Prendinger H, Ishizuka M (eds) Life-like characters. Tools, affective functions, and applications. Springer, Berlin, pp 463–468

Schabes Y (1990) Mathematical and computational aspects of lexicalized grammars. Ph.D. thesis, Computer Science Department, University of Pennsylvania

Scherer K (2000) Emotion. In: Hewstone M, Stroebe W (eds) Introduction to social psychology: a European perspective. Wiley-Blackwell, Oxford, UK, pp 151–191

Schröder M (2004) Speech and emotion research: an overview of research frameworks and a dimensional approach to emotional speech synthesis (Ph.D thesis). vol 7 of Phonus, Research Report of the Institute of Phonetics, Saarland University

Schröder M, Trouvain J (2003) The German text-to-speech synthesis system MARY: a tool for research, development and teaching. Int J Speech Tech 6: 365–377

Schröder M, Cowie R, Douglas-Cowie E, Westerdijk M, Gielen S (2001) Acoustic correlates of emotion dimensions in view of speech synthesis. In: Proceedings of Eurospeech 2001, Aalborg, Denmark, 3–7 September 2001, pp 87–90

Schröder M, Pirker H, Lamolle, Burkhardt F, Peter C, Zovato E (2010) Representing emotions and related states in technological systems. In: this volume

Searle J R (1969) Speech acts: an essay in the philosophy of language. Cambridge University Press, Cambridge

Serenari M, Dybkjaer L, Heid U, Kipp M, Reithinger N (2002) Survey of existing gesture, facial expression, and cross-modality coding schemes. IST-2000-26095 Deliverable D2.1, Project NITE

Stokoe WC (1978) Sign language structure: an outline of the communicative systems of the American deaf. Linstock Press, Silver Spring

Stone M, Bleam T, Doran C, Palmer M (2000) Lexicalized grammar and the description of motion events. In: TAG+5, Workshop on tree-adjoining grammar and related formalisms, Paris, France, 25–27 May 2000

Thiébaux M, Marsella S, Marshall AN, Kallmann M (2000) SmartBody: behavior realization for embodied conversational agents. In: Padgham L, Parkes DC, Müller J, Parsons S (eds) Proceedings of conference on autonomous agents and multi-agent systems (AAMAS08), Estoril, Portugal, 12–16 May 2008, pp 151–158

Trippel T, Gibbon D, Thies A, Milde JT, Looks K, Hell B, Gut U (2004) CoGesT: a formal transcription system for conversational gesture. In: Proceedings of 4th international conference on language resources and evaluation (LREC 2004), Lisbon, Portugal, 26–28 May 2004

Tsutsui T, Saeyor S, Ishizuka M (2000) MPML: a multimodal presentation markup language with character agent control functions. In: Proceedings of world conference on the WWW and internet, WebNet 2000, San Antonio, TX, USA, October 30–November 4

van Deemter K, Krenn B, Piwek P, Klesen M, Schröder M, Baumann S (2008) Fully generated scripted dialogue for embodied agents. Artif Intell J 172(10):1219–1244

Vilhjálmsson H, Cantelmo N, Cassell J, Chafai NE, Kipp M, Kopp S, Mancini M, Marsella S, Marshall AN, Pelachaud C, Ruttkay Z, Thórisson KR, van Welbergen H, van der Werf RJ (2007) The behavior markup language: recent developments and challenges. In: Pelachaud C et al (eds) Intelligent virtual agents. Springer, Berlin, pp 99–111

Walker M, Cahn J, Whittaker S (1996) Linguistic style improvisation for lifelike computer characters. In: Proceedings of the AAAI Workshop on AI, Alife and Entertainment. August, Portland, Oregon, USA

Wiggins J (1996) The five-factor model of personality: theoretical perspectives. The Guilford Press, New York, NY

Yue B, de Byl P (2006) The state of the art in game AI standardisation. In: Proceedings of the 2006 international conference on game research and development. December 4, Perth, Australia, ACM International conference proceedings series Vol 223. Murdoch University, Australia, pp 41–46

Zong Y, Dohi H, Ishizuka M (2000) Multimodal presentation markup language MPML with emotion expression functions attached. In: Proceedings of the international symposium on multimedia software engineering (IEEE Computer Soc), Taipei, Taiwan