Towards Experimental Specification and Evaluation of Lifelike Multimodal Behavior

BUISINE Stéphanie[1], ABRILIAN Sarkis[1], RENDU Christophe[1], MARTIN Jean-Claude[1&2]
martin@limsi.fr   http://www.limsi.fr/Individu/martin/research/projects/lea/

(1) LIMSI-CNRS, BP 133, 91403 Orsay Cedex, France
(2) LINC-Univ. Paris 8, IUT de Montreuil, 140 Rue de la Nouvelle France, 93100 Montreuil, France

Abstract

In this paper we introduce the Limsi Embodied Agent project which tackles the following issues of Embodied Conversational Agent (ECA) specification and evaluation: the need to ground ECA's behavior on video-taped annotations of application dependent human behavior, the granularity of the language for specifying the ECA multimodal behavior, and the evaluation of the use of ECA in the Human-Computer Interaction. In this paper, we describe preliminary work and future directions in each of these issues.

Introduction

There is still a lack of appropriate and global answers to the question of the "natural" behavior of Embodied Conversational Agent (ECA). The specification of multimodal behavior of ECA is often based on knowledge extracted from the literature in several domains such as Psychology, Sociology and Linguistics. As partly suggested by (Kipp 2001 ; Cassell et al. 2001b), we believe that in order to be lifelike, multimodal behavior of agents needs to be grounded in experimental studies in the same application context (i.e. the multimodal behavior of pedagogical ECA should be based on video recording and annotation of teacher's behavior in "similar" settings). In this paper, we describe how we intend to use such an experimental approach with the Limsi Embodied Agent (LEA). But how do we go from annotating human multimodal behavior to specifying the behavior of an ECA? Existing specification languages are mostly dedicated either to low-level monomodal specification (i.e. angry facial expression) or to amodal "higher" level specifications which are translated into monomodal features (i.e. angry behavior generating facial expression, intonation, gaze…). In the LEA project, we define an intermediate level of specification based on types of cooperation between communicative modalities which can be useful for fine-grain specification and evaluation of multimodal communicative behavior based on video corpus annotation (Martin et al. 2001).

Finally, we describe our global methodological framework which can be considered as a checklist for defining the evaluation process of ECAs.

1. Grounding ECA lifelike multimodal behavior on video annotation

Annotating human multimodal behavior

Following previous work on manual annotation of video-taped human multimodal behavior, we have developed tools making easier the annotation and the computation of behavioral metrics. We have defined a grammar for annotations (a XML DTD). According to this grammar, these annotations are composed of several sections. A first section describe the features the subject is referring to in the corpus (ie. objects drawn on the blackboard in the case of a teacher). Each of the following sections contains the annotation of a multimodal segment, itself composed of several sub-section (one for each modality such as speech, hand gesture, gaze) potentially including annotation of references to objects in each modality.

Computing metrics of human multimodal behavior

A Java software has been developed in order to parse these annotations of human multimodal behavior and to compute behavioral metrics (Martin et al. 2001). It follows the following steps:
- Parse the file containing the annotation and build internal representation
- Assign a « salience » value to each (object, reference) couple according to rules such as « *if the referent contains the fully specified name of the object, assign value 1.0 to the salience value* »
- Assign a priori fixed values to weights for each modality

- Compute the average salience value in each single multimodal segment across all modalities
- Compute the average salience value for each object across all modalities
- Compute behavioral metrics (complementarity/redundancy rate, equivalence rate)

From human behavior annotation to ECA behavior specification

Both the DTD and the software have been already applied to 40 samples taken in several corpora.
We are currently studying how to integrate this approach with the Anvil tool (Martin & Kipp 2002). We intend to evaluate such tools on larger corpora and to integrate them in a larger methodology for the analysis of multimodal behavior that we will apply to several domains such as e-learning (Martin et al. 2002). One long term goal we have is to find an efficient way for establishing a systematic mapping between annotations of human behavior and specifications of the multimodal behavior of the corresponding ECA. The resulting behavioral metrics (redundancy/complementarity rate, equivalence rate…) will form the basis of the language we propose for specifying the multimodal behavior of ECA that we describe in the next section.

2. Specifying cooperation between modalities in ECA behavior

Granularity level of existing ECA specification languages

Existing ECA specification languages can be compared on the basis of several criteria including the available modalities and the granularity of the specification tags (Table 1).
The VHML language (Gustavsson et al. 2001) is used to facilitate the interactions between a virtual agent and the user by featuring one specification sub-language for each modality (GML for gestures, SML for speech, BAML for body, FAML for facial expression) but also specification sub-languages for "higher" amodal levels (EML for emotion, DMML for Dialogue Manager Markup Language).
The BEAT project (Cassell et al. 2001c) enables the animation of an avatar by using typed text. It makes use of behavioral "suggestive functions", for example the "Surprising Feature Iconic Gesture Generator" function (movements generated when the avatar encounters surprising information). Behavior selection is achieved by two filters: one for the resolution of conflicts and the other for the priority threshold. BEAT is used in MACK (Cassell et al. 2002) which annotates automatically a text with the following modalities: hand gesture, gaze, eyebrow, body movement and intonation. In the specification language of REA (Cassell et al. 1999) different high-level functions combine several modalities. For example the "Give turn" function trains the hands' relaxation, a glance towards the user and the lifting of the eyebrows. The "Open interaction" function trains the eye to look towards the user, a smile and a head toss.
In (Traum & Rickel 2001) the language is specified to manage interactions between a group of immerged agents in a virtual world, and of which the specifications are in the form of conversational tags: make-contact, break-contact, give-attention, release-attention, start-topic, end-topic. Contrary to the works of Guerrin and al. (Guerrin et al. 2001) whose rules are related to a specific communication plan between two agents (seller and client), containing tags of a low level (defining specific expressions) and at the same time tags of a high level, which corresponds to combinations of low-level tags, or to works done by (Silva et al. 2001) where a unique active agent, the storyteller, is interacting with a passive user, and of which the specification language is based on four variables: behavior, environment, emotion and time of the day.

	Combi-nation of modalities	Conver-sational Behavior	Enviro-nment	Gaze	Arms & hands gestures	Head gestures	Face	Body posture	Speech		
									Recog-nizer	Synth-esizer	Intona-tion control
REA	+	++	-	+	+	+	+	+	+	+	+
BEAT	+	+	-	+	+	-	+	-	+	-	+
VHML	++	+++	-	+	-	+	+	-	-	+	+
Traum	+	++	+++	+	+	+	+	+	+	+	+
Guerrin	++	+	+	-	+	-	+	+	-	+	+
Storyteller	+	+	++	-	+	+	+	+	-	+	++

Table 1: Features of ECA specification languages (-: not used ; +++,++,+ : modality more or less used)

Low level specification of monomodal behavior

The current version of our LEA agent is written in Java and parses an XML file containing a sequence of configurations (Table 2). The current version is thus limited to the manual specification of each single configuration. The program uses single frame animation (gaze, facial expression, arms, head, body) and speech synthesis using IBMViaVoice and JavaSpeech API (Figure 1). A screendump is given in Figure 2.

```xml
<?xml version='1.0' encoding='utf-8'?>
<configurationsequence nbrconfig="2">
      <configuration>
            <timecode>1</timecode>
            <body>body.gif</body>
            <head>head-front.gif</head>
            <eyes>eyes-open-happy.gif</eyes>
            <gaze>pupils-middle.gif</gaze>
            <facialExpression>lips-open.gif</facialExpression>
            <bothArms>null</bothArms>
            <leftArm>arm-left-hello1.gif</leftArm>
            <rightArm>arm-right-hip.gif</rightArm>
            <speech>Hello, my name is LEA!</speech>
      </configuration>
      <configuration>
      …
      </configuration>
</configurationsequence>
```

Table 2: Low-level specification of each modality in the LEA agent. Each configuration specification features the image to be displayed for each body part.
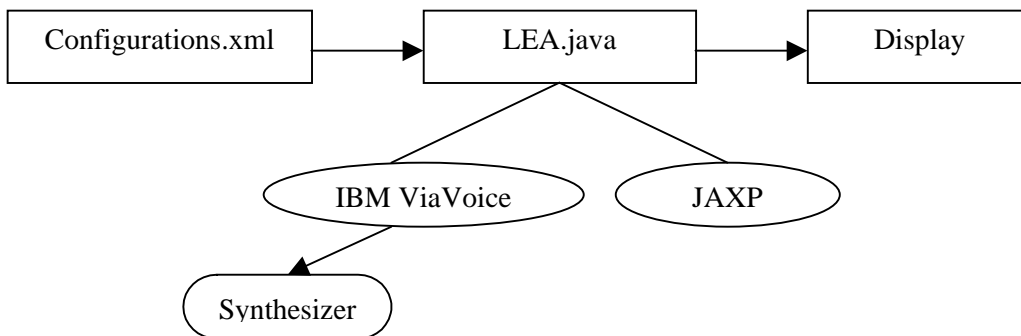

Figure 1: Current architecture of the LEA agent. The XML file containing a sequence of configuration specification is parsed by the LEA Java software with the JAXP API. Multimodal behavior is displayed via gif images and speech output using IBMViaVoice.

Figure 2: Screendump of the LEA agent[1].

---

[1] The graphical design of the LEA agent has been achieved by Christophe RENDU who can be joined at crendu@nexusanimation.com and +33.6.03.60.43.62

Towards "intermediate" specification of cooperation between modalities

We intend to augment the current specification of the LEA agent with "intermediate" level specification tags. This "intermediate" level of specification will be defined between the currently used low level of specification (ie. sequence of images) and a higher level of specification (ie. semantic representations, pragmatics and communicative goals…). This intermediate level of specification will be based on the Tycoon typology of cooperations between modalities (Martin et al. 2001). We believe that since this typology seems useful for the annotation of human multimodal behavior, it might also be useful to exhibit "natural" multimodal properties in ECA behavior:

- *Equivalence*: A cooperation by equivalence is defined by a set of modalities, a set of chunks of information, which can be displayed on either of the modalities and a criterion, which can be used by the agent to select one of the modalities. When several modalities cooperate by equivalence, this means that a chunk of information may be displayed as an alternative, by either of them.
- *Redundancy*: Several modalities, a set of chunks of information and two functions define a cooperation by redundancy. The first function can be used to find out the common attributes in chunks to be presented by the different modalities, the second function is used as a fission criterion. If modalities cooperate by redundancy, this means that these modalities will present the same information (ie. The values of several attributes of displayed monomodal information will overlap).
- *Complementarity*: Cooperation by complementarity is similar to cooperation by redundancy except that there are several non-common attributes between the chunks to be displayed by the different modalities.
- *Specialization* : Cooperation by specialization is defined by a modality, a set of modalities *A* and a set of chunks of information this modality is specialized in when compared to the modalities of the set *A*. When modalities cooperate by specialization, this means that a specific kind of information is always displayed by a single modality.
- *Concurrency* : Cooperation by concurrency means that several modalities display independent chunks of information at the same time.

Such intermediate tags specifying cooperation between modalities might be integrated with the low-level tags in different ways. One possibility is to use one file `TycoonAgent.xml` defining the multimodal behavior (or personality) of the agent thanks to Tycoon tags (ie. equivalence/redundancy…), and a second file `Configurations.xml` containing the initial presentation that the avatar must achieve. The specifications provided in `TycoonAgent.xml` would then act as a filter of the presentation specified in `Configurations.xml` in order to extract the effective multimodal expressions of LEA (Figure 3).
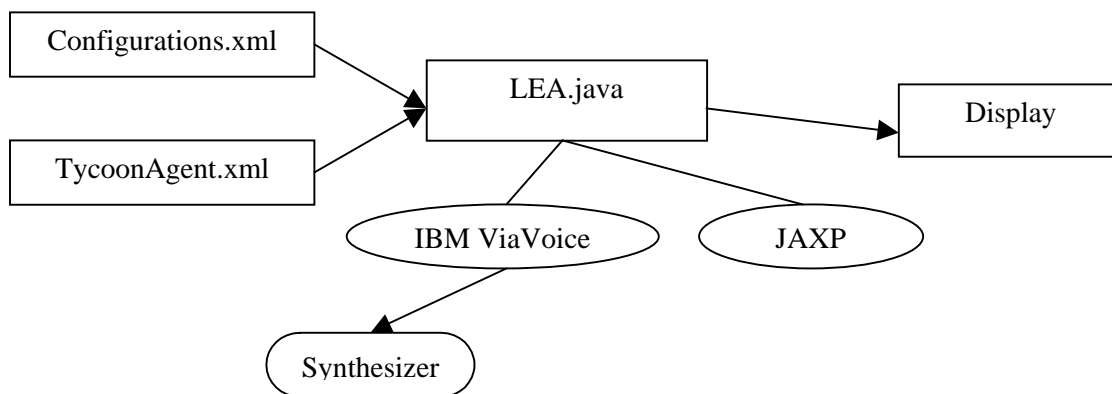

Figure 3: Independent specification of the ECA's multimodal personality (TycoonAgent.xml) and the sequence of multimodal configurations (Configurations.xml).

Another possibility is to also include Tycoon tags in the configurations file. It would then be possible to make for example the agent more or less redundant at certain times. Example: < **redundancy**>right</**redundancy** > will lead the avatar to do gesture, body turning and gaze towards right. Cascaded multimodal style sheet might be used: Tycoon tags provided in "Configurations.xml" would have priority. If there is not any, those of "TycoonAgent.xml" would be used as default multimodal behavior.

3. Methodological issues regarding the evaluation of ECA

Grounding the specification of ECA multimodal behavior in data and grammar resulting of human multimodal behavior annotation does not ensure that the ECA system will improve the interaction with the user when compared with non ECA systems. In this third and last section, we describe a global methodological framework which can be considered as a checklist for defining the evaluation process of ECAs.

Conversation with a believable embodied agent must include (Cassell et al., 1994) verbal communication and visual nonverbal behaviors enriching the communication, such as facial movements and hand gesture. In the field of Human Computer Interaction (HCI), the opportunity to use spontaneous speech and gesture communication, close to the one humans use with one another, might be a way of improving both the effectiveness and the pleasantness of the interaction. However, this expected effect remains hypothetical and needs to be validated, as well as guidelines to achieve a human-like interaction remain to be clearly defined. In this research domain, the prevailing method for collecting preliminary data is the 'Wizard of Oz' technique: the experimenter, hidden from the user, controls the agent's behavior and simulates an 'intelligent' system capable of understanding and responding to the user's spontaneous speech and gesture. Besides providing a way of observing the user's behavior, this method could allow sharp evaluations of embodied conversational agents. In this context, we think that there is still a need of specifying valid ways of testing theoretical hypotheses. In present time, since Psychology is more and more emphasized in ergonomics training, HCI evaluations begin to be conducted within the methodological framework of experimental psychology. The main features of this methodology will be briefly reviewed in this section. After clarifying the hypotheses we focus on, we list variables worth considering in the evaluation. Some notions important to take into account in the experimental design of ECA experimental studies are then defined, as well as a few statistical principles.

Testable hypotheses

We propose in Table 3 three hypotheses that should lead the evaluation of ECAs. We would like to point out that an hypothesis must be based on the assumption of a difference. Hypotheses concerning an absence of difference (null hypotheses) are not statistically testable. Although quite intuitive, the hypotheses proposed below need to be further examined. For example, previous research concerning H2 failed to show any difference in performance when using interfaces with or without an agent (André et al., 1998). Moreover, from the fact that embodied agent systems differ from classical interfaces in several factors (verbal communication, visual nonverbal cues) and that they differ among themselves in several features (choices of conception, degree of sophistication…), it is important to identify relevant variables and to dissociate them in the test of the hypotheses. As an example, higher perceived helpfulness of the system and enhanced engagement and entertainment have been attributed to the inclusion of an agent in the interface (see Cassell et al., 2001 for a review). However, as far as we know, the use of speech and nonverbal communication in input and/or in output were not crossed in tnese experiments. Thus, their relative influence and their relationship (interactive, additive, etc.) have not been tested. Therefore, variables contributing to the specificity of embodied agents systems are listed in the following section.

	Hypothesis	To-be-measured criteria
H1	The use of a conversational agent enhances the ergonomics of the interface.	(Bastien & Scapin, 1993) criteria: guidance; workload; explicit control; adaptability; error management; consistency; significance of codes; compatibility.
H2	The use of a conversational agent enhances the effectiveness of the interaction.	speed of interaction (time to achieve the goal, navigation); performance or achievement of intermediate goals (number of errors).
H3	The use of a conversational agent enhances the satisfaction of the user.	self-rated pleasantness, effectiveness, usefulness, ease of use, ease to learn...

Table 3: The proposed list of hypotheses to guide the evaluation of embodied agent systems.

Variables to manipulate

A variable is defined as a situational or an individual characteristic having a possible influence, according to the experimenter's hypotheses, on the studied situation (Matalon, 1969). To be testable, a variable must comprise several values (at least two: presence vs. absence of the characteristic). We enumerate in Table 4 variables (and their respective values) interesting to test in the evaluation of ECA systems. Although several variables can be crossed in factorial designs, the whole list cannot be tested in a single experiment whatever. The reader should rather consider it as a check-list (as exhaustive as possible) of potentially contributing factors to the usefulness of embodied agent systems.

	Variables	Values	Remarks
V1	use of verbal communication	in input; in output; both; none	
V2	additional vocal cues	true / false	e.g. intonation, intensity, pitch, tone…
V3	additional nonverbal communication	in input; in output; both; none	
V4	type of nonverbal cues	gesture; facial expression; both; none	
V5	type of multimodal cooperation	equivalence; specialization; transfer; redundancy; complementarity; concurrency; none	TYCOON (Martin et al., 2001)
V6	agent's abilities and skills	ability to engage a conversation; personality; emotions; none	several values can coexist within a single agent.
V7	amount of embodiment	face-only; full-body	
V8	style of rendering	2D; 3D	
V9	Realism	cartoonist; photo-realistic	
V10	sophistication of animation	still images; cartoon; realistic	
V11	type of voice in output	synthetic; natural	
V12	fit of the agent with the user's characteristics	true / false	e.g. same age, same sex...
V13	fit of the agent with the user's preferences	true / false	e.g. opposite sex, funny characters for children...
V14	fit of the agent with the task characteristics	true / false	e.g. air hostess for a flight reservation service...

Table 4: The proposed list of variables to manipulate in the evaluation of ECA systems.

Some empirical results have already been obtained with independent tests of some of these variables. Indeed, the use of speech in both input and output (V1) without any embodied agent (speech-based system), has proved to be preferred and rated as better than navigation in a classical interface (Caelen & Bruandet, 2001). Could additional nonverbal communication (V3) improve this effect? Besides the fact that subjects may feel awkward speaking loudly without the face of a dialogue partner (Seto et al., 1994), visual nonverbal communication seems to have an additive influence on the effectiveness of the interaction (Granström et al., 2002), to enhance the probability that the user understands the agent's speech and its emotional state (Massaro et al., 2000) and to shorten the total time in speech dialogue (Seto et al., 1994). However, in a pedagogical context, Moreno et al. (2001) demonstrated that the visual presence of an agent (either a fictional one or a video of a human face) does not affect performance. Concerning the agent's skills (V6), the possibility to have a social dialogue proved to enhance trust in the service for a certain category of users (Bickmore & Cassell, 2001). Moreover, the expression of emotion (position of eyebrows and mouth corner) seems to be useful in the understanding of speech (Massaro et al., 2000). On the other hand, the realism (V9) is not of prime importance in the effectiveness of the interaction (Vilhjalmsson and Cassell, 1998). Finally, Kohar and Ginn (1997) recommend the use of dramatized agents instead of fully developed human personalities (V12), arguing that they make better interface agents.

These empirical results are useful in that they orient the conception of new embodied agent systems. Most of them arose from rigorous experimental designs, but, with certain exceptions, they did not resort to statistical analyses likely to allow their generalization. We propose, after a brief recall of experimental principles, to review the main advantages and constraints of statistical methods.

Experimental design

The values of the variables retained for the experiment determine the experimental groups. For example, the value corresponding to the absence of a factor define the so-called control group. On the contrary, all the groups must be equivalent regarding the values of all non-tested variables. Actually, the purpose of an experimental design is to put into evidence the influence of the manipulated variables, but also to exclude the influence of potential interfering variables (Matalon, 1969). Thus, the experimenter must, as a preliminary, identify as many interfering variables as possible. In the HCI domain, the strongest one is probably the skill-level towards use of computers. This variable must then be controlled: either all the subjects present the same level, or the diversity of levels is the same within all groups. Another interfering variable can be the order of the conditions each subject performs: to neutralize its effect, this order must be counterbalanced. Finally, as far as possible, the groups must be of the same size, so as to enhance the sensitivity of statistical comparisons.

Statistical analysis

Whereas notions directing the construction of experimental designs go a great deal into HCI research, principles of statistical analyses and their usefulness seem to remain mostly unknown. As experimental research is aimed at testing theoretical assumptions within a population of subjects or objects (Le Ny and Gineste, 1995), statistics provide a means of collecting data within only a sample of users and then testing if the obtained results can be generalized to the whole population. A representative sample is composed of subjects possessing the same characteristics as the parent population (age, sex, social group, level of education...) and selected from it at random. Existing statistical methods can be classified into three categories: descriptive methods, parametric inferential methods and nonparametric inferential methods. Descriptive statistics are used for summing up and organizing the data. Only inferential statistics, which are based on probabilistic theories, allow generalizing the obtained results to the whole population. Whereas parametric inferential methods are quite restricting in the constitution of the subjects sample (size, adjustment to a probability law...), nonparametric (distribution free) methods are more flexible.

4. Conclusion

We proposed here a methodological framework taking into account experimental psychology and statistics principles in the field of HCI, and especially embodied agent systems specification and evaluation. We currently intend to apply this methodology for the evaluation of the LEA agent in several application domains such as home environment, education and games.

5. References

André, E., Rist, T., Muller, J. (1998). Integrating reactive and scripted behaviors in a life-like presentation agent. Proceedings of AGENTS'98, pp; 261-268.

Bastien, J.M.C., Scapin, D.L. (1993). Ergonomic criteria for the evaluation of Human-Computer Interfaces. INRIA Technical Report, 156.

Bickmore, T., Cassell, J. (2001). A relational agent: a model and implementation of building user trust. Proceedings of the CHI'01 Conference, pp. 396-403. March 31-April 5, Seattle, Washington.

Caelen, J., Bruandet, M.F. (2001). Interaction multimodale pour la recherche d'information. In: C. Kolski (Ed), Environnements évolués et évaluation de l'IHM, pp. 175-205. Paris: Hermès Science Publications.

Cassell, J., Bickmore, T., Billinghurst, M., Campbell, L., Chang, K., Vilhjálmsson, H. and Yan, H. (1999). "Embodiment in Conversational Interfaces: Rea." Proceedings of the CHI'99 Conference, pp. 520-527. Pittsburgh, PA.

Cassell, J., Bickmore, T., Campbell, L., Vilhjalmsson, H., Yan, H. (2001a). More than just a pretty face: conversational protocols and the affordances of embodiment. Knowledge-Based Systems, 14, 55-64.

Cassell, J., Nakano, Y., Bickmore, T., Sidner, C., Rich, C. (2001b) Non-Verbal Cues for Discourse Structure. Proceedings of the 41st Annual Meeting of the Association of Computational Linguistics, pp. 106-115. July 17-19, Toulouse, France

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. Proceedings of SIGGRAPH '94, pp. 413-420.

Cassell, J., Stocky, T., Bickmore, T., Gao, Y., Nakano, Y., Ryokai, K., Tversky, D., Vaucelle, C., Vilhjálmsson, H. (2002) MACK: Media lab Autonomous Conversational Kiosk. Proceedings of Imagina02. February 12-15, Monte Carlo.

Cassell, J., Vilhjálmsson, H., Bickmore, T.(2001c) BEAT: the Behavior Expression Animation Toolkit. Proceedings of SIGGRAPH '01, pp. 477-486. August 12-17, Los Angeles, CA.

Granström, B., House, D., Swerts, M. (2002). Multimodal feedback cues in human-machine interactions. In: Proceedings of the Speech Prosody 2002 Conference, 11-13 April, Aix-en-Provence, pp. 347-350.

Guerrin, F., Kamyab, K., Arafa, Y., Mamdani, E. (2001) Conversational Sales Assistants. Proceedings of the workshop on " Representing, Annotating, and Evaluating Non-Verbal and Verbal Communicative Acts to Achieve Contextual Embodied Agents ", May 29, 2001, Montreal, in conjunction with The Fifth International Conference on Autonomous Agents. pp 35-40.

Gustavsson, C., Strindlund, L., Wiknertz, E., Beard, S., Huynh, Q., Marriott, A., Stallo, J. (2001). Virtual Human Markup Language. http://www.vhml.org/.

Kipp, M. (2001) Analyzing Individual Nonverbal Behavior for Synthetic Character Animation" In: C. Cave, I. Guaitella, S. Santi (eds.) Oralité et Gestualité - Actes du colloque ORAGE 2001, Paris: L'Harmattan, pages 240-244.

Kohar, H., Ginn, I. (1997). CHI 97: Mediators: guides through online TV services. CHI 97 Electronic Publications.

Le Ny, J.F., Gineste, M.D. (1995). Démarches et méthodes. In: J.F.Le Ny and M.D. Gineste (Eds.), La Psychologie ; Textes Essentiels, pp. 15-18. Paris: Larousse.

Martin, J.C. & Kipp, M. (2002). Annotating and Measuring Multimodal Behaviour - Tycoon Metrics in the Anvil Tool. Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC'2002), Las Palmas, Canary Islands, Spain, 29-31 may 2002

Martin, J.C. , Réty, J.H., Bensimon, N. (2002). Multimodal and Adaptative Pedagogical Resources. Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC'2002), Las Palmas, Canary Islands, Spain, 29-31 may 2002 http://www.lrec-conf.org/lrec2002/index.html

Martin, JC., Grimard, S., Alexandri, K. (2001). On the annotation of the multimodal behavior and computation of cooperation between modalities. Proc. of the workshop on "Representing, Annotating, and Evaluating Non-Verbal and Verbal Communicative Acts to Achieve Contextual Embodied Agents", May 29, 2001, Montréal, in conjunction with The 5th Int. Conf. on Autonomous Agents. pp 1-7.

Massaro, D.W., Cohen, M.M., Beskov, J., Cole, R.A. (2000). Developing and evaluating conversational agents. In: Cassell, J., Sullivan, J., Prevost, S., Churchill, E. (Eds.). Embodied conversational agents, pp. 287-318. MIT Press.

Matalon, B. (1969). La logique des plans d'expérience. In: G. Lemaine, J.M. Lemaine (Eds.), Psychologie Sociale et Expérimentation. Paris: Mouton/Bordas.

Maybury, M. & Martin, J.-C. (2002) Proceedings of the workshop "Multimodal Resources and Multimodal Systems Evaluation". M. Maybury & J.-C. Martin (Eds). Third international conference on language resources and evaluation (LREC'2002), Las Palmas, Canary Islands, Spain, 1st June 2002

Moreno, R., Mayer, R.E., Spires, H.A., Lester, J.C. (2001). The case for social agency in computer-based teaching: do students learn more deeply when they interact with animated pedagogical agents? Cognition and Instruction, 19, 177-213.

Sanders, G.A., Scholtz, J. (2000). Measurement and evaluation of embodied conversational agents. In: J. Cassell, J. Sullivan, S. Prevost, E. Churchill (Eds.). Embodied conversational agents, pp 346-373, MIT Press.

Seto, S., Kanazawa, H., Shinshi, H., Takebayashi, Y. (1994) Spontaneous speech dialogue system TOSBURG II and its evaluation. Speech Communication, 15, 341-353.

Silva, A., Vala, M., Paiva, A. (2001). The Storyteller : Building a Synthetic Character That Tells Stories. Proceedings of the workshop on "Representing, Annotating, and Evaluating Non-Verbal and Verbal Communicative Acts to Achieve Contextual Embodied Agents", May 29, 2001, Montreal, in conjunction with The Fifth International Conference on Autonomous Agents. pp 53-58.

Traum, D., Rickel, J. (2001) Embodied Agents for Multi-party Dialogue in Immersive Virtual Worlds. Proceedings of the workshop on " Representing, Annotating, and Evaluating Non-Verbal and Verbal Communicative Acts to Achieve Contextual Embodied Agents ", May 29, 2001, Montreal, in conjunction with The Fifth International Conference on Autonomous Agents. pp 27-33.

Vilhjalmsson, H.H., Cassell, J. (1998). Bodychat: autonomous communicative behaviors in avatars. ACM Proceedings of the 2nd International Conference on Autonomous Agents, Minneapolis, may 9-13.