The Acorformed Corpus: Investigating Multimodality in Human-Human and Human-Virtual Patient Interactions

M. Ochs[1,2], P. Blache[1,3], G. Montcheuil[1,3,5], J.M. Pergandi[1,4], R. Bertrand[1,3], J. Saubesty[1,3], D. Francon[6], and D. Mestre[1,4]

[1]Aix Marseille Université, Université de Toulon, CNRS,
[2]LIS UMR 7020, [3]LPL UMR 7309, [4]ISM UMR 7287 ; [5]Boréal Innovation,
[6]Institut Paoli-Calmettes (IPC), Marseille, France
magalie.ochs@lis-lab.fr, blache@lpl-aix.fr

Abstract

The paper aims at presenting the Acorformed corpus composed of human-human and human-machine interactions in French in the specific context of training doctors to break bad news to patients. In the context of human-human interaction, an audiovisual corpus of interactions between doctors and actors playing the role of patients during real training sessions in French medical institutions have been collected and annotated. This corpus has been exploited to develop a platform to train doctors to break bad news with a virtual patient. The platform has been exploited to collect a corpus of human-virtual patient interactions annotated semi-automatically and collected in different virtual reality environments with different degree of immersion (PC, virtual reality headset and virtual reality room).

1 Introduction

For several years, there has been a growing interest in Embodied Conversational Agents (ECAs) to be used as a new type of human-machine interface. ECAs are autonomous entities, able to communicate verbally and nonverbally (Cassell, 2000). Indeed, several researches have shown that embodied conversational agents are perceived as social entities leading users to show behaviors that would be expected in human-human interactions (Krämer, 2008).

Moreover, recent research has shown that virtual agents could help human beings *improve their social skills* (Anderson et al., 2013; Finkelstein et al., 2013). For instance in (Anderson et al., 2013), an ECA endowed the role of a virtual recruiter is used to train young adults to job interview. In our project, we aim at developing a virtual patient to train doctors to break bad news. Many works have shown that doctors should be trained not only to perform medical or surgical acts but also to develop skills in communication with patients (Baile et al., 2000; Monden et al., 2016; Rosenbaum et al., 2004). Indeed, the way doctors deliver bad news has a significant impact on the therapeutic process: disease evolution, adherence with treatment recommendations, litigation possibilities (Andrade et al., 2010). However, both experienced clinicians and medical students consider this task as difficult, daunting, and stressful. Training health care professional to break bad news is now recommended by several national agencies (e.g. the French National Authority for Health, HAS)[1].

A key element to exploit embodied conversational agents for social training with users is their *believability* in terms of socio-emotional responses and global multimodal behavior. Several research works have shown that non-adapted behavior may significantly deteriorate the interaction and the learning (Beale and Creed, 2009). One methodology to construct believable virtual agent is to develop a model based on the analysis of a corpus of human-human interaction in the social training context (as for instance in (Chollet et al., 2017)). In our project, in order to create a virtual patient with believable multimodal reactions when the doctors break bad news, we have collected, annotated, and analyzed two multimodal corpora of interaction in French in this context. Both human-human and human-machine interaction are considered to investigate the effects of the virtual reality displays on the interaction. In this paper, we present the two corpora in the following sections.

[1]The French National Authority for Health is an independent public scientific authority with an overall mission of contributing to the regulation of the healthcare system by improving health quality and efficiency.

Magalie Ochs, Philippe Blache, Grégoire Montcheuil, Jean-Marie Pergandi, Roxane Bertrand, Jorane Saubesty, Daniel Francon and Daniel Mestre 2019. The Acorformed Coprus: Investigating Multimodality in Human-Human and Human-Virtual Patient Interactions. *Selected papers from the CLARIN Annual Conference 2018*. Linköping Electronic Conference Proceedings 159: 113–120.

2 Multimodal Human-Human Corpus Analysis to Model Virtual Patient's Behavior

The modeling of the virtual patient is based on an audiovisual corpus of interactions between doctors and actors playing the role of patients (called "Standardized patients") during real training sessions in French medical institutions (it is not possible, for ethical reasons, to record real breaking bad news situations). The use of "Standardized Patients" in medical training is a common practice. The actors are carefully trained (in our project, actors are also nurses) and follow pre-determined scenarios defined by experts to play the most frequently observed patients reactions. The recommendations of the experts, doctors specialized in breaking bad news situations, are global and related to the attitude of the patient ; the verbal and non-verbal behavior of the actor remains spontaneous. Note that the videos of the corpus have been selected by the experts as representative of real breaking bad news situations.

On average, a simulated consultation lasts 9 minutes. The collected corpus, in French, is composed of 13 videos of patient-doctor interaction (the doctor or the patient vary in the video), with different scenarios[2].

The initial corpus has been semi-manually annotated, leading to a total duration of 119 minutes. Different tools have been used in order to annotate the corpus. First, the corpus has been automatically segmented using SPPAS (Bigi, 2012) and manually transcribed using Praat (Boersma, 2002). The doctors' and patient's non-verbal behaviors have been manually annotated using ELAN (Sloetjes and Wittenburg, 2008). Different gestures of both doctors and patients have been annotated: head movements, posture changes, gaze direction, eyebrow expressions, hand gestures, and smiles. Three annotators coded the corpus. Each of them annotated a third of the corpus. The annotators were graduate students in linguistics and were paid to annotate. In order to insure homogeneity among the annotators, a guide was given describing every annotation steps. Moreover, the annotations' sessions were supervised, allowing the annotators to ask questions at any moment. In order to validate the annotation, 5% of the corpus has been annotated by one more annotator. The inter-annotator agreement, using Cohen's Kappa, was satisfying (k=0.63). More details on the corpus are presented in (Porhet et al., 2017).

2.1 Verbal cues annotations

Audio files were extracted from the video recordings. The speech signal was segmented into Inter-Pausal Units (IPUs), defined as speech blocks surrounded by at least 200 ms silent pauses. Due to its objective nature (Koiso et al., 1998), the IPU can be automatically segmented. However, due to poor audio quality, they were manually corrected. We manually transcribed each participant's speech on two different tiers using the TOE convention (Transcription Orthographique Enrichie / Enriched Orthographical Transcription, (Bertrand et al., 2008)). Note that we do not consider the acoustic features (e.g. prosody) since the audio quality of the videos does not enable us to study this aspect. The part-of-speech (POS) tags were automatically identified using MarsaTag (Rauzy et al., 2014). MarsaTag is a stochastic parser for written French which has been adapted to account for the specificities of spoken French. Among other outputs, it provides a morpho-syntactic category for each POS token.

2.2 Visual cues annotations

Different modalities of both the doctors and the patients have been annotated. The modalities as well as the corresponding values are described in Table 1[3] .

We summarize the annotation for each interlocutor in Table 2. The table reveals that the most frequent non-verbal signals are the doctor's and patient's head movements while few smiles appear. The number of words shows that the doctors speak more than the patient, as expected given the context of the interaction.

[2]The corpus is on Ortolang part of the CLARIN infrastructure

[3]As we are interested only in movements, we did not differentiated one movement from another. The hand annotation indicate the time interval from the moment the hands start moving until they return to the rest position.

Modality	Values
Head movements	nod, shake (negation), tilt, bottom, up, side
Posture change (movements of the bust)	forward, backwards, other change
Gaze direction	oneself, interlocutor, other direction, closed eyes
Eyebrow expression	frown, raise
Hand gesture	movement
Smile	smile, no smile

Table 1: Non-verbal modalities

Category	Doctors	Patients
Head	3649	1970
Hands	635	463
Gaze	1823	716
Smile	20	20
Eyebrows	225	189
Posture	239	257
Words	44816	727

Table 2: Total number of annotations per interlocutor

The annotated corpus has been analyzed for three different purposes:

- to build the *dialog model of the virtual patient*: the dialog model of the virtual patient is based on the notion of "*common ground*" (Garrod and Pickering, 2004; Stalnaker, 2002), *i.e.* a situation model represented through different variables that is updated depending on the information exchange between the interlocutors. The variables describing the situation model (e.g. the cause of the damage), specific to breaking bad news situations, have been defined based on the manual analysis of the transcribed corpus and in light of the pedagogical objective in terms of dialog. The dialog model is described in more detail in (Ochs et al., 2017) ;
- to design *non-verbal behaviors of the virtual patient*: the corpus has been used to enrich the non-verbal behavior library of the virtual patient with gestures specific to breaking bad news situations.
- to design *the feedback behavior of the virtual patient*: in order to identify the multimodal signals triggering feedback from the patients, we have applied sequences mining algorithms to extract rules to model the multimodal feedback behavior of the virtual patient (for more details (Porhet et al., 2017)).

3 Multimodal Human-Virtual Patient Corpus Analysis to Investigate the Users' experience with different virtual reality displays

Based on the corpus analysis presented in the previous section, we have implemented a virtual reality training system inhabited by a virtual patient and developed to give the capabilities to doctors to simulate breaking bad news situation. The system is *semi-autonomous* since it includes both automatic and manual modules, making it possible to simulate a fully automatized human-machine interaction (for more details on the semi-autonomous system (Ochs et al., 2018a)). Implemented on three different virtual environment displays (PC, virtual reality headset, and an immersive virtual reality room), the doctors can interact in natural language with a virtual patient that communicates through its verbal and non-verbal behavior (Figure 1). In order to collect the interaction and create the corpus of human-machine interaction in the context of breaking bad news, we have implemented a specific methodology. First, the doctor is filmed using a camera. His gestures and head movements are digitally recorded from the tracking data: his head (stereo glasses), elbows and wrists are equipped with tracked targets. A high-end microphone synchronously records the participant's verbal expression. As for the virtual agent, its gesture and verbal

Figure 1: Participants interacting with the virtual patient with different virtual environment displays (from left to right): virtual reality headset, virtual reality room, and PC.

expressions are recorded from the Unity Player. The visualization of the interaction, is done through a 3D video playback player we have developed (Figure 2). This player replays synchronously the animation and verbal expression of the virtual agent as well as the movements and video of the participant.

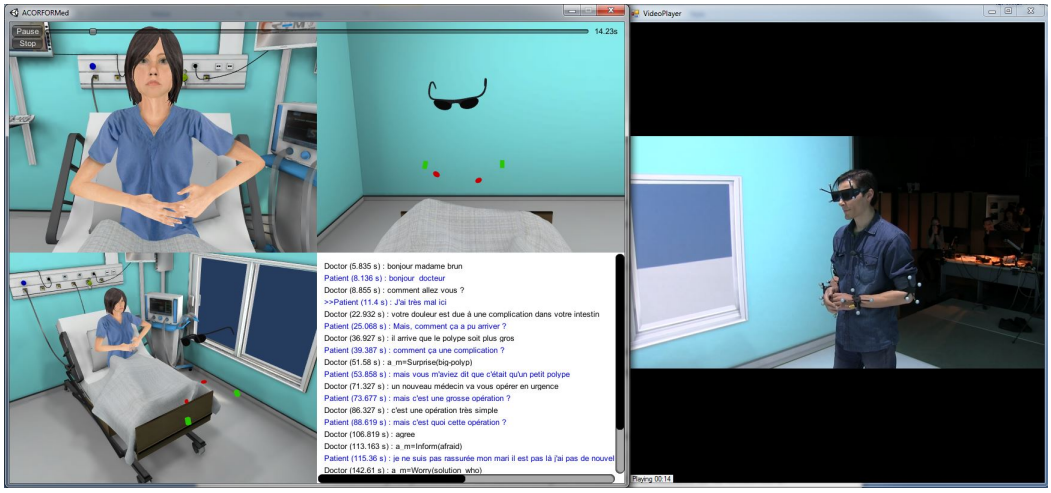

Figure 2: 3D video playback player

This environment facilitates the collection of corpora of doctor-virtual patient interaction in order to analyze the verbal and non-verbal behavior in different immersive environments.

Using the semi-autonomous system, we have collected 108 interactions in French of participants with the virtual patient. In total, 36 persons have participated in the experimentation. Ten of them are real doctors that already have experience in breaking bad news to real patients. Each participant has interacted with the systems 3 times with three different devices: PC, virtual reality headset, and virtual reality room. The task of the participants was to announce a digestive perforation after a gastroenterologic endoscopy in immediate post operative period[4]. The collected corpus is composed of 108 videos (36 per device). The total duration of the corpus is 5h34 (among which two hours with real doctors). In average, an interaction lasts 3mn16 (an example of interaction is presented on the following web page http://www2.lpl-aix.fr/ acorformed/videos.html).

3.1 Segmentation

The interaction between the participants and the virtual patient is split into 3 phases: the beginning, the central part, and the conclusion. Based on a previous analysis of human-human interaction in the same context (Saubesty and Tellier, 2016), we suppose that the verbal and nonverbal behavior may differ de-

[4]The scenario has been carefully chosen with the medical partners of the project for several reasons (e.g. the panel of resulting damages, the difficulty of the announcement, its standard characteristics of announce).

pending on the phases of the interaction. Keeping this in mind, we performed our analysis independently for each phase for all the data sources we have. We defined the size of each phase relative to the total duration of the interaction. As a first step, we define empirically the duration of each phase: 15% of the total conversation for the introduction, 70% for the central part of the interaction, and 15% for the conclusion. Note that a script in Python has been written with the percentage of each phase in parameter to automatically compute the verbal and non-verbal cues described in the following with different segmentation.

3.2 Verbal cues

In order to analyze the verbal behavior of the participants, we have defined high-level characteristics reflecting the *lexical richness* and the *linguistic complexity* of the user's verbal behavior based on the frequency of the part-of-speech tags for each participant and each phase of the interaction. Using a specific tool called SPPAS (Bigi, 2012), we performed a tokenization followed by a phonetization on the transcription file. The part-of-speech (POS) tags were automatically identified using MarsaTag (Rauzy et al., 2014). We consider 9 parts-of-speech tags: adjective, adverb, auxiliary, conjunction, determiner, noun, preposition, pronoun, verb.

Based on these POS tags, we computed the *lexical richness*, measured as the fraction of adjectives and adverbs out of the total number of tokens and the *linguistic complexity*, measured as the fraction of conjunctions, prepositions and pronouns out of the total number of token. The descriptive statistics are reported Table 3.

	Introduction		Central part		Conclusion	
	Average	SD	Average	SD	Average	SD
Lexical Richness	0.16	0.07	0.15	0.03	0.18	0.09
Linguistic Complexity	0.15	0.05	0.17	0.03	0.19	0.08
Length of the sentences	6.24	4.04	8.73	1.70	7.86	3.73
Length of IPUs	1.92	1.97	1.92	0.40	2.76	3.32

Table 3: Average and Standard Deviation (SD) of the verbal cues per phase.

Moreover, we have computed the *length of the sentences in terms of number of words* and the *lengths of inter-pausal units in terms of duration*. We compute the average length of sentences in each phase of the interaction for each participant. The length corresponds to the number of words of a sentence. The MarsaTag tool (Rauzy et al., 2014) has been used to define the sentences from the transcript text. The speech signal was segmented into Inter-Pausal Units (IPUs), defined as speech blocks surrounded by at least 200 ms silent pauses[5]. Due to its objective nature, the IPU has been automatically segmented using SPASS (Bigi, 2012). The descriptive statistics are reported Table 3.

3.3 Non-Verbal cues

Concerning the non-verbal cues, we have computed the *entropy* to characterize the movements of the participant in the virtual environment. The entropy is a common measure in virtual reality domain to assess the movements of the participants (Maïano et al., 2011). To obtain the entropy of the curve defined by the movement of each tracker on the participant, following the method described in (Dodson et al., 2013), we have computed the upper-bound on the Shannon entropy of curves of each plane (x, y and z) and each tracked point (head, left wrist, right wrist, left elbow, and right elbow). Finally, the different computed values of entropy are averaged to obtained two non-verbal cues: the average movements of the head, and the average movement of the arms. The descriptive statistics are reported Table 4.

[5]For French language, lowering this 200 ms threshold would lead to many more errors due to the confusion of pause with the closure part of unvoiced consonants, or with constrictives produced with a very low energy.

	Introduction		Central part		Conclusion	
	Average	SD	Average	SD	Average	SD
Head	1.61	0.43	2.94	0.55	1.55	0.45
Arms	1.31	0.42	2.47	0.44	1.31	0.49

Table 4: Average and Standard Deviation (SD) of the non-verbal cues per phase.

3.4 Sense of presence

In order to evaluate the global experience of the users, we asked the participants to fill different questionnaires on their subjective experience to measure their feeling of presence (with the *Igroup Presence Questionnaire*, IPQ (Schubert, 2003)), feeling of co-presence (Bailenson et al., 2005), and perception of the believability of the virtual patient (questions extracted from (Gerhard et al., 2001))[6]. These subjective evaluations enabled us to *tag* the video of the corpus with the results of these tests and then to correlate objective measures (e.g. verbal and non-verbal cues of the participants) to subjective measures (e.g. feeling of presence and perception of the virtual patient's believability) using machine learning methods (for more details see (Ochs et al., 2018b)).

4 Conclusion

In this article, we have presented two multimodal *comparable* corpora. The corpora have been collected in the same context of doctors' trainings to break bad news but in two different conditions of interaction: human-actor patient and human-virtual patient. They have been analyzed manually and using data mining methods in order to construct an autonomous virtual reality training platform inhabited by a virtual patient.

Given the different natures of the corpora, different annotations techniques - manual, semi-automatic and automatic - have been used, leading to different annotations schemes. Our next step is to harmonize the annotations of the two corpora in order to compare the verbal and non-verbal behaviors of the doctors depending on the type of the interaction. Our final goal is to identify objective verbal and non-verbal cues that could reflect the engagement of the user in the interaction with the virtual patient based on the verbal and the non-verbal cues identified in the human-human interaction.

Acknowledgements

This work has been funded by the French National Research Agency project ACORFORMED (ANR-14-CE24-0034-02) and supported by grants ANR-16-CONV-0002 (ILCB), ANR-11-LABX-0036 (BLRI) and ANR-11-IDEX-0001-02 (A*MIDEX).

5 Bibliographical References

References

K. Anderson, E. André, T. Baur, S. Bernardini, M. Chollet, E. Chryssafidou, I. Damian, C. Ennis, A. Egges, P. Gebhard, et al. 2013. The tardis framework: intelligent virtual agents for social coaching in job interviews. In *Advances in computer entertainment*, pages 476–491. Springer.

A. D. Andrade, A. Bagri, K. Zaw, B. A. Roos, and Ruiz J. G. 2010. Avatar-mediated training in the delivery of bad news in a virtual world. *Journal of palliative medicine*, 13(12):1415–1419.

W. Baile, R. Buckman, R. Lenzi, G. Glober, E. Beale, and A. Kudelka. 2000. Spikes—a six-step protocol for delivering bad news: application to the patient with cancer. *Oncologist*, 5(4):302–311.

J. N. Bailenson, C. Swinth, K.nd Hoyt, S. Persky, A. Dimov, and J. Blascovich. 2005. The independent and interactive effects of embodied-agent appearance and behavior on self-report, cognitive, and behavioral markers of copresence in immersive virtual environments. *Presence: Teleoperators and Virtual Environments*, 14(4):379–393.

[6]The analyze of the subjective experience of the participants is out of scope of this paper and is described in an other article (Ochs et al., 2018c)

R. Beale and C. Creed. 2009. Affective interaction: How emotional agents affect users. *International journal of human-computer studies*, 67(9):755–776.

Roxane Bertrand, Philippe Blache, Robert Espesser, Gaëlle Ferré, Christine Meunier, Béatrice Priego-Valverde, and Stéphane Rauzy. 2008. Le cid-corpus of interactional data-annotation et exploitation multimodale de parole conversationnelle. *Traitement automatique des langues*, 49(3):pp–105.

B. Bigi. 2012. Sppas: a tool for the phonetic segmentations of speech. In *The eighth international conference on Language Resources and Evaluation*.

P. Boersma. 2002. Praat, a system for doing phonetics by computer. *Glot international*, 13(341-345).

J. Cassell. 2000. More than just another pretty face: Embodied conversational interface agents. *Communications of the ACM*, 43:70–78.

M. Chollet, M. Ochs, and C. Pelachaud. 2017. A methodology for the automatic extraction and generation of non-verbal signals sequences conveying interpersonal attitudes. *IEEE Transactions on Affective Computing*.

Michael Maurice Dodson, Michel Mendes France, and Michel Mendes. 2013. On the entropy of curves. *Journal of Integer Sequences*, 16(2):3.

S. Finkelstein, S. Yarzebinski, C. Vaughn, A. Ogan, and J. Cassell. 2013. The effects of culturally congruent educational technologies on student achievement. In *International Conference on Artificial Intelligence in Education*, pages 493–502. Springer.

S. Garrod and M. Pickering. 2004. Why is conversation so easy? *Trends in cognitive sciences*, 8(1):8–11.

M. Gerhard, D. J Moore, and D. Hobbs. 2001. Continuous presence in collaborative virtual environments: Towards a hybrid avatar-agent model for user representation. In *International Workshop on Intelligent Virtual Agents*, pages 137–155. Springer.

Hanae Koiso, Yasuo Horiuchi, Syun Tutiya, Akira Ichikawa, and Yasuharu Den. 1998. An analysis of turn-taking and backchannels based on prosodic and syntactic features in japanese map task dialogs. *Language and speech*, 41(3-4):295–321.

N. Krämer. 2008. Social effects of virtual assistants. a review of empirical results with regard to communication. In *Proceedings of the international conference on Intelligent Virtual Agents (IVA)*, pages 507–508, Berlin, Heidelberg. Springer-Verlag.

Christophe Maïano, Pierre Therme, and Daniel Mestre. 2011. Affective, anxiety and behavioral effects of an aversive stimulation during a simulated navigation task within a virtual environment: A pilot study. *Computers in Human Behavior*, 27(1):169–175.

K. Monden, L. Gentry, and T. Cox. 2016. Delivering bad news to patients. *Proceedings (Baylor University. Medical Center)*, 29(1).

M. Ochs, G. Montcheuil, J-M Pergandi, J. Saubesty, B. Donval, C. Pelachaud, D. Mestre, and P. Blache. 2017. An architecture of virtual patient simulation platform to train doctor to break bad news. In *International Conference on Computer Animation and Social Agents (CASA)*.

M. Ochs, P. Blache, G. Montcheuil, J.-M. Pergandi, J. Saubesty, D. Francon, and D. Mestre. 2018a. A semi-autonomous system for creating a human-machine interaction corpus in virtual reality: Application to the acorformed system for training doctors to break bad news. In *Proceedings of LREC*.

Magalie Ochs, Sameer Jain, Jean-Marie Pergandi, and Philippe Blache. 2018b. Toward an automatic prediction of the sense of presence in virtual reality environment. In *Proceedings of the 6th International Conference on Human-Agent Interaction*, pages 161–166. ACM.

Magalie Ochs, Daniel Mestre, Grégoire De Montcheuil, Jean-Marie Pergandi, Jorane Saubesty, Evelyne Lombardo, Daniel Francon, and Philippe Blache. 2018c. Training doctors' social skills to break bad news: evaluation of the impact of virtual environment displays on the sense of presence. *Journal on Multimodal User Interfaces*, pages 1–11.

C. Porhet, M. Ochs, J. Saubesty, G. Montcheuil, and R. Bertrand. 2017. Mining a multimodal corpus of doctor's training for virtual patient's feedbacks. In *Proceedings of 19th ACM International Conference on Multimodal Interaction (ICMI), Glasgow, UK*.

Stéphane Rauzy, Grégoire Montcheuil, and Philippe Blache. 2014. Marsatag, a tagger for french written texts and speech transcriptions. In *Proceedings of Second Asian Pacific Corpus linguistics Conference*, page 220.

M. Rosenbaum, K. Ferguson, and J. Lobas. 2004. Teaching medical students and residents skills for delivering bad news: A review of strategies. *Acad Med*, 79.

J. Saubesty and M. Tellier. 2016. Multimodal analysis of hand gesture back-channel feedback. In *Gesture and Speech in Interaction, Nantes, France*.

T. Schubert. 2003. The sense of presence in virtual environments: A three-component scale measuring spatial presence, involvement, and realness. *Zeitschrift für Medienpsychologie*, 15(69-71).

H. Sloetjes and P. Wittenburg. 2008. Annotation by category: Elan and iso dcr. In *6th International Conference on Language Resources and Evaluation*.

R. Stalnaker. 2002. Common ground. *Linguistics and Philosophy*, 25(5):701–721.