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*

M. Ochs[1], P. Blache[2], G. Montcheuil[2,3,4], J.-M. Pergandi[3], J. Saubesty[2], D. Francon[5], and D. Mestre[3]
Aix Marseille Université CNRS ENSAM, Université de Toulon, [1]LSIS UMR 7296,
[2]LPL UMR 7309, [3]ISM UMR7287 ; [4]Boréal Innovation, [5]Institut Paoli-Calmettes (IPC), Marseille, France

Abstract

In this paper, we introduce a two-step corpora-based methodology, starting from a corpus of human-human interactions to construct a semi-autonomous system in order to collect a new corpus of human-machine interaction, a step before the development of a fully autonomous system constructed based on the analysis of the collected corpora. The presented methodology is illustrated in the context of a virtual reality training platform for doctors breaking bad news.

Keywords: Virtual patient, Virtual reality, Corpus, Health domain, Training

1. Introduction

This paper presents a *corpus-based methodology* elaborated to design a *virtual reality environment* aiming at training doctors to *break bad news to a virtual patient.* 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). Among all possible bad new, doctors can be faced to the complex situation of announcing a damage associated to a care, and that can engage their responsibility: unforeseeable medical situation, dysfunction, medical error, etc. The way *doctors deliver bad news related to damage associated with care* 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]. Such trainings are organized as workshops during which doctors disclose bad news to actors playing the role of patients. This solution is complex to implement: it requires several persons, it is costly, and time consuming (each 30 mn. session requires one hour of preparation). Our project[2] aims at developing *a virtual reality training system* with an embodied conversational agent playing the role of a virtual patient.

The approaches classically used to develop an interactive system including an embodied conversational agent consist in analyzing automatically or manually *one* annotated corpus of human-human interactions (Cassell, 2000). Then, a fully autonomous prototype is developed and evaluated through perceptive studies. In the approach presented here, the methodology is based on several steps, making it possible to acquire specific data and information about the required modules and functionalities. More precisely, the procedure starts classically with the collection of natural data (recordings of human-human interactions), but we introduce a new step based to acquire human-machine interactional data. The idea consists in developing a conversational agent implementing both automatic and manual modules, making it possible to simulate a fully automatized human-machine interaction, without needing to develop the entire system (in particular the comprehension module).

This methodology presents several advantages. First, both human-human and human-machine interaction corpora correspond to the same context. Their study leads to specify the verbal or non-verbal behavioral characteristics that the trainee may use when faced with a virtual human compared to interpersonal interactions. As a consequence, the prototype completely fits with the observed users behavior during human-machine interaction. Moreover, using a semi-autonomous system enables one to abstract from critical modules, difficult to develop and crucial for a successful interaction, for example the speech recognition system, that may strongly deteriorate the interaction in case of failure. Such modules are integrated in a third step without any impact on the evaluation of the first prototype. In the remaining of this paper, we illustrate an application of this methodology with the development of a prototype for training doctors to break bad news to patients.

The paper is organized as follows. In the first section, we present the global methodology based on different corpora. In section 3, we present the multimodal corpus of human-human interaction used to develop the first semi-autonomous prototype. Section 4 is dedicated to the presentation of the semi-autonomous platform and the different tools developed to collect an automatically annotated corpus of human-machine interaction. In section 5, we describe the human-machine corpus and how it is used to develop the final prototype.

[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.

[2] ACORFORMed, (Ochs et al., 2017), http://www.lpl-aix.fr/ãcorformed/.

2. The Corpus-based Methodology

The methodology used for developing the virtual training platform consists of several steps, each based on the creation and the analysis of a corpus (human-human or human-machine interaction) in the context of breaking bad news situations. These steps can be specified as follows (see Figure 1):

1. Analysis of a human-human interaction corpus: specification of the discourse organization, definition of the different behavioral modules, development of a prototype (a semi-autonomous agent).
2. The prototype is used to collect a new corpus of human-machine interaction.
3. The new human-machine corpus is analyzed to specify the missing modules and improve the different functionalities in order to develop the autonomous version of the system.
4. Evaluation of the system, classically with perceptive experiments.

The initial corpus of doctor-patient interaction (recorded during training sessions) has been annotated manually (see section 3). The results of the analysis have been used to develop a first version of the virtual reality training platform. This platform, as described in Section 4, is semi-autonomous: some modules of the architecture are simulated by an experimenter. This semi-autonomous platform has been used to collect a new corpus of human-machine interaction, considering different devices of interaction (PC, virtual reality headset, and virtual reality cave). The collected corpora constitute the basis of the missing modules and improve the initial prototype in the perspective of a fully autonomous virtual training platform.

3. Multimodal Human-Human Corpus Analysis to Model Virtual Agent's Behavior

The modeling of the virtual patient is based on an audio-visual 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 is composed of 13 videos of patient-doctor interaction (each video involves a different doctor and/or a different actor-patient pair), with different scenarios[3]. The size of the corpus remains small. However, our objective is not to learn a model (in a machine learning point of view) but to extract automatically and manually information to model the virtual patient's behavior (Porhet et al., 2017) that, then, will be validated through perceptive studies.

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 and Weenik, 1996). The doctors' and patients 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 experts annotated one third of the corpus each. 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). The corpus and the annotations are described in more detail in (Porhet et al., 2017).

The annotated corpus has been analyzed for two 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. We used for the implementation the dialog system OpenDial (Lison and Kennington, 2016). In this approach, the model selects automatically the verbal patient's reaction depending on the recognized verbal utterance of the doctor (matched to flexible patterns defined through regular expressions) as well as the history of the dialog (the information already delivered by the doctor being encoded in the *common ground*). 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. The recurrent gestures identified in the corpus are those used to indicate pain. In total, we have enriched the virtual character's non-verbal behavior library with 16 new gestures specific to this context.

[3] The corpus is on Ortolang

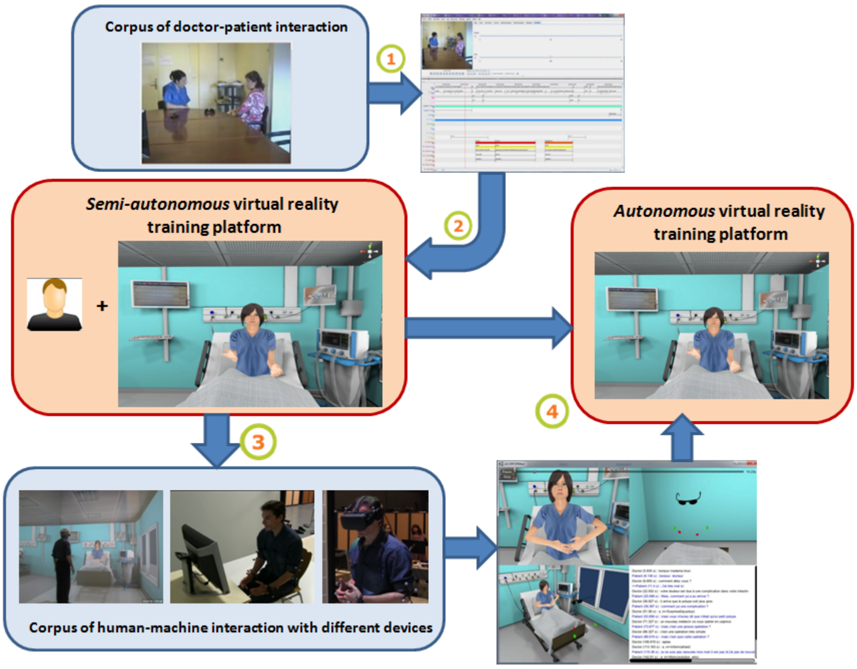

Corpus of doctor-patient interaction
1
2
Semi-autonomous virtual reality training platform
+
Autonomous virtual reality training platform
3
4
Corpus of human-machine interaction with different devices

Figure 1: Methodology

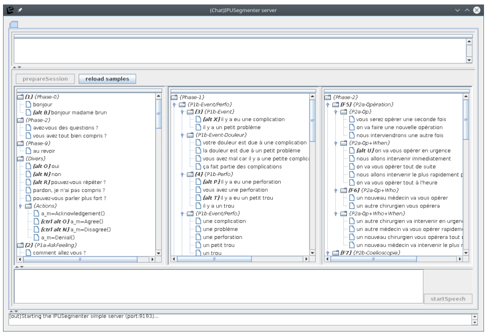
The dialog model as well as the non-verbal behavior library of the virtual patient have been used to implement a first prototype of a virtual reality training platform, as described in the next section.

4. Semi-Autonomous Virtual Reality Training Platform

In order to evaluate the general architecture of the training platform and to collect a corpus of human-machine interaction specific to our project, we have developed a semi-autonomous platform, which architecture is described in Figure 2.

The platform is *semi-autonomous* because some modules of the system are automatic (for example the dialogue generation) where some others are manual. In particular, the speech recognition and the comprehension modules are simulated by a human: the doctor verbal production is interpreted in real time by the operator which selects the adequate input signal to be transmitted to the dialogue system. Indeed, these modules may be particularly critical in case of failure and then damage strongly the interaction. They represent moreover the most difficult part of the system to be developed. Replacing the module by the operator comes to a perfect speech recognition and comprehension. This makes it possible to completely control the corresponding parameters and concentrate on the evaluation of the others modules, such as the dialog supervision and the non-verbal behavior of the virtual patient. Moreover, it renders possible the evaluation of the overall interaction (e.g. presence, satisfaction, believability).

A specific interface has been designed for this purpose to enable the experimenter to select the sentences as close as possible to that has been said by the doctors (Figure 3).



[1] {Phase-0}
bonjour
[alt B] bonjour madame brun
{Phase-2}
avez-vous des questions ?
vous avez tout bien compris ?
{Phase-9}
au revoir
{Divers}
[alt O] oui
[alt N] non
[alt R] pouvez-vous répéter ?
pardon, je n'ai pas compris ?
pouvez-vous parler plus fort ?
{Actions}
a_m=Acknowledgement()
[ctrl alt O] a_m=Agree()
[ctrl alt N] a_m=Disagree()
a_m=Denial()
[2] {P1a-AskFeeling}
comment allez vous ?

{Phase-1}
{P1b-Event/Perfo}
[3] {P1b-Event}
[alt X] il y a eu une complication
il y a un petit problème
{P1b-Event-Douleur}
votre douleur est due à une complication
la douleur est due à un petit problème
vous avez mal car il y a une petite complica
ça fait partie des complications
[4] {P1b-Perfo}
[alt P] il y a eu une perforation
vous avez une perforation
[alt T] il y a eu un petit trou
il y a un trou
{P1b-Event/Perfo}
une complication
une problème
une perforation
un petit trou
un trou

{Phase-2}
[F5] {P2a-Opération}
{P2a-Op}
vous serez opérer une seconde fois
on va faire une nouvelle opération
nous interviendrons une autre fois
{P2a-Op+When}
[alt U] on va vous opérer en urgence
nous allons intervenir immediatement
on va vous opérer tout de suite
nous allons intervenir le plus rapidement p
on va vous opérer tout à l'heure
[F6] {P2a-Op+Who}
un nouveau médecin va vous opérer
un autre chirurgien vous opérera
{P2a-Op+Who+When}
un autre chirurgien va intervenir en urgen
un autre médecin va vous opérer rapidem
un nouveau chirurgien vous opérera tout
un nouveau médecin va intervenir le plus
[F7] {P2b-Coelioscopie}

[out]Starting the IPUSegmenter simple server (port:9193)...

Figure 3: Screen-shot of the interface of the experimenter to select the corresponding doctor's recognized sentences

The interface contains 136 prototypical sentences (or patterns) organized into different dialog phases: greetings, asking the patient's feelings, description of the surgical problem, description of the remediation. These sentences have been defined based on the analysis of the transcribed corpus of doctor-patient interaction (see section 3). Each prototypical sentence encodes a family of possible utterances, as identified in the corpus. The sentences are encoded into an XML file. Keyboard shortcuts are associated to each sentence/pattern, and can be configured in order to be easily selected by the experimenter. Several pre-tests have been built to test the interface and train the experimenter. Note that at the difference with a "Wizard of Oz", the experimenter does not select the virtual patient's reaction but only send to the dialog model the recognized doctor's sentence. In fact in general, in a Wizard of Oz setup, the experimenter plays the role of both the recognition module and the dialog module (understanding and selection of the response). In the proposed set-up, only the recogni-

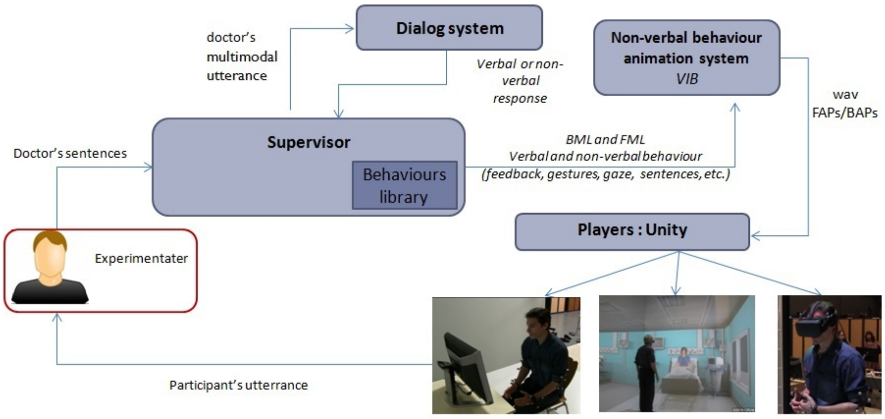

Dialog system
doctor's multimodal utterance
Verbal or non-verbal response
Non-verbal behaviour animation system VIB
wav FAPs/BAPs
Doctor's sentences
Supervisor
Behaviours library
BML and FML
Verbal and non-verbal behaviour
(feedback, gestures, gaze, sentences, etc.)
Players : Unity
Experimentater
Participant's utterrance

Figure 2: Overall architecture of the virtual reality training platform

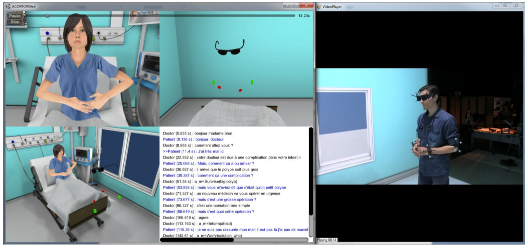
tion module is replaced by the experimenter.

The dialogue system then generates a sequence of instructions, to be sent to a non-verbal behavior animation system called VIB (Pelachaud, 2009). This system computes the animation parameters (*Facial Animation Parameters* - FAP - and *Behavioral Animation Parameters* - BAP) to animate the face and the body of the virtual patient. The result is encoded in XML and describes the communicative intention to perform (encoded in FML, *Function Markup Language*) as well as the non-verbal signals to express (encoded in BML, *Behavior Markup Language*). Moreover, the VIB system contains a text-to-speech synthesis (Aylett and Pidcock, 2007) for generating the speech, in synchronization with the non-verbal behavior (including lips animation).

In order to experiment as broadly as possible the validity of the approach, we have implemented the virtual patient on different platforms: PC, virtual reality headset, and virtual reality cave. The virtual reality cave is constituted of a 3m deep, 3m wide, and 4m high cubic space with three vertical screens and a horizontal screen (floor). A cluster of graphics machine makes it possible to deliver stereoscopic, wide-field, real-time rendering of 3D environments, including spatial sound. This offers an optimal sensorial immersion of the user. The environment has been designed to simulate a real recovery room where the breaking bad news are generally performed. The virtual agent based on the VIB platform has been integrated in by means of the *Unity* player.

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 expressions are recorded from the Unity Player. The visualization of the interaction, is done through a 3D video playback player we have developed (Figure 4). This player replays synchronously the animation and verbal expression of the virtual agent as well as the movements and video of the participant.

Figure 4: 3D video playback player



Doctor (5.835 s) : bonjour madame brun
Patient (8.136 s) : bonjour docteur
Doctor (8.855 s) : comment allez vous ?
>>Patient (11.4 s) : J'ai très mal ici
Doctor (22.932 s) : votre douleur est due à une complication dans votre intestin
Patient (25.068 s) : Mais, comment ça a pu arriver ?
Doctor (36.927 s) : il arrive que le polype soit plus gros
Patient (39.387 s) : comment ça une complication ?
Doctor (51.58 s) : a_m=Surprise(big-polyp)
Patient (53.858 s) : mais vous m'aviez dit que c'était qu'un petit polype
Doctor (71.327 s) : un nouveau médecin va vous opérer en urgence
Patient (73.677 s) : mais c'est une grosse opération ?
Doctor (86.327 s) : c'est une opération très simple
Patient (88.619 s) : mais c'est quoi cette opération ?
Doctor (106.819 s) : agree
Doctor (113.163 s) : a_m=Inform(afraid)
Patient (115.36 s) : je ne suis pas rassurée mon mari il est pas là j'ai pas de nouvel
Doctor (142.61 s) : a_m=Worry(solution_who)

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. The collected corpus is the basis of the development of a fully autonomous virtual reality training platform.

5. From Human-Machine Corpus to a Fully Autonomous Training Platform

Using the semi-autonomous system described in the previous section, we have collected 108 interactions of participants with the virtual patient. In total, 36 persons have participated to the experimentation. Ten of them are real doctors that already have an experience in breaking bad news to real patients. The others are student from the University. 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]. Before the interaction, written instructions were presented to the participants: the role they have to play is a doctor that had just operated the virtual patient to remove a polyp in the bowel. A digestive perforation occurred during the surgery.

[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).

These written instructions explains precisely the causes of the problem, the effects (pain), and the proposed remediation (a new surgery, urgently). Participants are asked to read the instructions several times as well as before each interaction. The understanding is verified by means of a questionnaire. Each participant has the instruction to announce this medical situation to the virtual patient three times with three different devices: PC, virtual reality headset, and virtual reality room. The order of the conditions were counterbalanced.

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 ACORFORMed site). Note that thanks to the tools described in the previous section, some of the non-verbal participant behavior can be automatically annotated.

In order to validate this first prototype, 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)). 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 behavior of the participants) to subjective measures (e.g. feeling of presence and perception of the virtual patient's believability).

We are currently analyzing the corpus before entering into the development of the fully autonomous training platform (in particular the comprehension and generation module). We already have used the corpus to train and test the speech recognition system, in order to ensure that the speech recognition system can accurately recognize the participants. We also verify that the recognized words and sentences activate correctly the expected rules in the dialog model. These comprehension rules are adapted in consequence. Moreover, the corpus is also used in order to compare the non-verbal behaviors through the different devices. We formulate the hypothesis that participants use less gestures in the virtual reality headset condition since they do not see their body. Depending on the results, the automatic gestures recognition could be adapted.

6. Conclusion

The development of embodied conversational agent with specific conversational skills is a challenge for human-machine communication, not only concerning the dialogue capacities of the agent, but also the adequacy of the environment to the task. The creation of rich and large corpora is a pre-requisite towards this goal. We have presented in this paper a semi-automatic platform answering these needs: offering the possibilities of an ecological human-machine interaction within a virtual reality environment following the final architecture of the system by implementing or simulating the functionalities of all the modules.

7. 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).

8. Bibliographical References

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

Aylett, M. P. and Pidcock, C. J. (2007). The cerevoice characterful speech synthesiser sdk. In *IVA*, pages 413–414.

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

Bailenson, J. N., Swinth, K., Hoyt, C., Persky, S., Dimov, A., and Blascovich, J. (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.

Bigi, B. (2012). Sppas: a tool for the phonetic segmentations of speech. In *The eighth international conference on Language Resources and Evaluation*, pages 1748–1755.

Boersma, P. and Weenik, D. (1996). Praat: a system for doing phonetics by computer. report of the institute of phonetic sciences of the university of amsterdam. *Amsterdam: University of Amsterdam.*

Cassell, J. (2000). *Embodied conversational agents*. MIT press.

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

Gerhard, M., Moore, D. J., and Hobbs, D. J. (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.

Lison, P. and Kennington, C. (2016). OpenDial: A Toolkit for Developing Spoken Dialogue Systems with Probabilistic Rules. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Demonstrations)*, pages 67–72, Berlin, Germany. Association for Computational Linguistics.

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

Ochs, M., Montcheuil, G., Pergandi, J.-M., Saubesty, J., Donval, B., Pelachaud, C., Mestre, D., and Blache, P. (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)*.

Pelachaud, C. (2009). Studies on gesture expressivity for a virtual agent. *Speech Communication*, 51(7):630–639.

Porhet, C., Ochs, M., Saubesty, J., Montcheuil, G., and Bertrand, R. (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.*

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

Schubert, T. W. (2003). The sense of presence in virtual environments: A three-component scale measuring spatial presence, involvement, and realness. *Zeitschrift für Medienpsychologie*, 15(2):69–71.

Sloetjes, H. and Wittenburg, P. (2008). Annotation by category: Elan and iso dcr. In *LREC*.

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