Simulating emotional reactions in medical dramas

Sandra Williams and **Richard Power** and **Paul Piwek** [1]

Abstract.

Presenting information on emotionally charged topics is a delicate task: if bare facts alone are conveyed, there is a risk of boring the audience, or coming across as cold and unfeeling; on the other hand, emotional presentation can be appropriate when carefully handled, but when overdone or mishandled risks being perceived as patronising or in poor taste. When Natural Language Generation (NLG) systems present emotionally charged information linguistically, by generating scripts for embodied agents, emotional/affective aspects cannot be ignored. It is important to ensure that viewers consider the presentation appropriate and sympathetic.

We are investigating the role of affect in communicating medical information in the context of an NLG system that generates short medical dramas enacted by embodied agents. The dramas have both an informational and an educational purpose in that they help patients review their medical histories whilst receiving explanations of less familiar medical terms and demonstrations of their usage. The dramas are also personalised since they are generated from the patients' own medical records. We view generation of natural/appropriate emotional language as a way to engage and maintain the viewers' attention. For our medical setting, we hypothesize that viewers will consider dialogues more natural when they have an enthusiastic and sympathetic emotional tone. Our second hypothesis proposes that such dialogues are also better for engaging the viewers' attention.

As well as describing our NLG system for generating natural emotional language in medical dialogue, we present a pilot study with which we investigate our two hypotheses. Our results were not quite as unequivocal as we had hoped. Firstly, our participants did notice whether a character sympathised with the patient and was enthusiastic. This did not, however, lead them to judge such a character as behaving more naturally or the dialogue as being more engaging. However, when pooling data from our two conditions, dialogues with versus dialogues without emotionally appropriate language use, we discovered, somewhat surprisingly, that participants did consider a dialogue more engaging if they believed that the characters showed sympathy towards the patient, were not cold and unfeeling, and were natural (true for the female agent only).

1 INTRODUCTION

Consider the following three extracts of interactions between a senior nurse and a junior (student) nurse in medical dramas generated by our system:

A **Senior:** Radiotherapy targets cancer cells.
Junior: Cool!

B **Senior:** Anaemia is a condition in which patients feel very tired and may become breathless.
Junior: Right.

C **Junior:** So, let's hope that the packed red cell transfusion took care of the anaemia.
Senior: Yes.

How might viewers perceive the junior nurse's reactions? To the answer in A, the junior responds enthusiastically, perhaps excited by the medical technology, whereas to the one in B, the junior responds more neutrally, perhaps indirectly showing awareness of the patient's discomfort. In C, the junior's summary could be perceived as sympathetic to the patient. Of course, the response in A and summary in C might be perceived as sarcastic and the response in B as unfeeling. If a more direct empathetic response had been attempted in B, e.g., "Oh dear!", or "That's bad!", then it might be perceived as more natural, but it could also be interpreted as patronising or unprofessional. Interestingly, if there were no response at all, the characters might also come across as cold and unfeeling, whilst an inappropriate enthusiastic response such as D might make the characters appear macabre:

D **Senior:** A radical mastectomy is an operation to remove the breast.
Junior: Cool!

We are exploring the simulation of emotions in such responses and their effect on viewers's perceptions of the attitudes of the embodied agents. Our hope is that by generating dialogues in which the characters produce language that is sympathetic to the viewer/patient and enthusiastic about medical technology where appropriate, this will lead to:

- viewers perceiving the dialogue as natural/appropriate;
- engaging the attention of the viewers.

The presentation of emotionally charged information is fraught with difficulties, particularly if the viewer is the patient whose medical record is being discussed (as is our ultimate aim). Our hypotheses connect specific ways of presenting medical information that take emotion into account with perceived naturalness of the resulting dialogues and also the extent to which the dialogues are engaging. The two hypotheses are linked by the underlying idea that appropriate emotional responses will make the dramas more engaging: the viewers' attention will be captured, forcing them to listen more carefully to the interchanges and soak up medical information in the process.

In this preliminary work, we limited our study to enthusiastic responses such as the one in A, neutral responses such as the one in B and sympathetic summaries such as "So, let's hope that the packed red cell transfusion took care of the anaemia". We modified our system to produce such responses in generated dialogue and conducted

[1] The Open University, Walton Hall, Milton Keynes, MK7 6AA, U.K., E-mail: s.h.williams@open.ac.uk, r.power@open.ac.uk, p.piwek@open.ac.uk

a pilot study to elicit viewers' perceptions of two conditions: (a) with emotional responses and summaries and (b) with no responses and neutral summaries (see the Appendix).

2 THE MEDICAL DRAMAS

Our generated medical dramas present a discussion between a senior and junior nurse about a patient's medical record (the system has access to a simulated repository of breast cancer patients' medical records). The senior nurse asks the junior to read the patient's notes for a particular date and, as he reads the notes, the junior nurse also asks questions about medical terms; the senior explains these terms and elaborates on the various medical investigations and interventions that the patient underwent. Consequently, our system generates a type of tutorial dialogue in which the senior nurse is tutor and the junior is student.

The main difference of our approach with other work on tutorial dialogue (e.g., [19]) is that we generate both sides of the conversation as a drama script, just as one might generate a linear text. The differences from generating monologue are that we need to simulate the kinds of questions, answers and explanations that would take place in a dialogue between a tutor and student. One advantage is that we can explore generation of the language of dialogue turns without any necessity for natural language understanding, which would be required in conventional natural language dialogue systems where only half of the conversation is machine-generated.

An obvious consequence is that the user is a viewer, not a participant in the dialogue or the drama. Since the viewer is one step removed she cannot pose her own questions to the system. This might appear a disadvantage on first consideration but it is actually an advantage, for two reasons. First, students rarely have the ability to ask good questions, although they can be taught how ([6]). The viewer can learn from watching the drama unfold, and one important motivation for presenting a tutorial dialogue drama is to demonstrate to viewers how to ask questions. Our aim to provide them with an experience from which they can learn vicariously not only the answers to the questions, but also how to ask questions of their own — a benefit of presentations in dialogue form that has been demonstrated in previous work (e.g., [4, 3]). Second, researchers have found that when people interact with screen characters, they have false expectations of human-like qualities which the characters cannot fulfil, and that sometimes characters can make them feel stupid (see [14]). There is thus a danger that an interactive experience could be frustrating or annoying, so we think our aims are better met by a presentation in which the patient views a video of characters interacting with each other.

Our first pilot experiment was with a version of our system in which medical information was presented as a bare sequence of question and answer dialogue turns with no reactions to the information being presented. Eleven participants listened to a dialogue and a monologue generated from the same underlying electronic health record; they answered some comprehension and preference questions and wrote comments [18]. There was no difference in comprehension or preferences, however; the main comment was that the medical information was too closely packed, so that people had difficulty following it. We came up with a number of solutions for spacing out the medical information and presenting it more slowly. The solution that we will highlight in this paper is that of adding affective reactions to the medical information (other solutions will be reported elsewhere).

Figure 1. Screen shot from an output video.

3 THE SYSTEM

Our NLG system is a data-to-dialogue system — that is, the input is data and the output is a script for a dialogue. It builds a dialogue by querying a simulated relational database of breast cancer patients' medical records; builds concept graphs from the query results (a fragment of a concept graph is shown in Figure 2); adds questions and dictionary definitions to the original concept graph (Figure 3); plans dialogue turns; and realises them as a script for an embodied agent drama. The script is then performed using Loquendo text-to-speech software and Cantoche LivingActor^{TM}character animation (a screen shot of the output is shown in Figure 1). The system is described in more detail in [18].

Figure 2. Part of a concept graph built from data retrieved from a database of medical records.

Figure 2 depicts two concepts, a medical intervention and a medical problem, linked by an arrow representing an INDICATED BY relation between them. The meaning can be paraphrased as "anaemia motivated a packed red cell transfusion". A content planner in the NLG system augments this structure by adding questions and definitions from the system's dictionary of medical terms. Figure 3 illustrates how these would be added to the fragment in Figure 2; the rectangles and arrow from Figure 2 are shown greyed-out and new rectangles representing questions, a definition from the dictionary, and an attribute of the definition, are shown in black.

3.1 Defining medical terms

Our planner adds explanations of medical terms only if they have not been mentioned previously in the dialogue, and only if they are relatively rare in everyday language. Our information on term frequencies was derived from searches of the British National Corpus, a 100 million word corpus of British English (www.natcorp.ox.ac.uk). Our

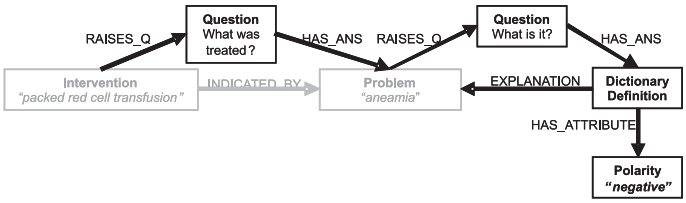

Figure 3. The graph augmented with questions and dictionary definitions.

searches revealed that medical terms such as "anaemia" and "axilla" are infrequent in the BNC with 362 and 3 occurrences, respectively, so these are defined, whereas "breast" was more frequent with 1,615 occurences, therefore it is not defined. However, BNC frequencies did not always coincide with our intuitions about whether people would know a term, for instance, "armpit" only occurs 76 times in the BNC, even though we believe that it is a well-known term. Consequently, we were guided by the BNC, but rather than following a rigid rule to define all terms within a fixed range of BNC frequencies (e.g., 0 to 1,000), we were also guided by our intuitions. In effect we took the medical terms that had a low frequency in the BNC and then selected a subset that we deemed suitable for explanation.

When the content planner adds an explanation of a medical term, it looks it up the definition in its medical term dictionary. The system's dictionary is a text file of definitions that we found on trusted Web sites such as www.cancerresearchuk.org; some fragments of the dictionary follow:

```
TERM anaemia
DEF NPS NEG a condition in which patients feel
 very tired and may become breathless
CAUSE S NEG the blood has fewer red blood
 cells than normal

TERM axilla
DEF NPS NEUTRAL the armpit

TERM CTScan
DEF NPS POS an X-ray scan using a computer
 to construct pictures of the body in cross
 section
```

Here, definitions for the terms "anaemia", "axilla" and "CTScan" are shown. The keyword TERM indicates the beginning of a new term and it is followed by a string containing the term. DEF and CAUSE indicate the beginning of a term's definition and cause (if any), NPS and S are syntactic categories (singular noun phrase and sentence), POS, NEG and NEUTRAL indicate the *polarity* of the definition or cause.

By polarity, we mean whether the definition or cause conveys information that is potentially beneficial, neutral, or detrimental from a patient's point of view. Remember that the medical records input to our system are simulated from patients who have breast cancer. Medical procedures such as radiation therapy or chemotherapy that destroy cancer cells are assigned positive polarity (as is CTScan in the fragment above). Obviously this is a somewhat naive view since although some medical technologies can potentially help patients, some also have unpleasant side effects. Negative polarities are assigned typically to definitions of illnesses, such as anaemia, which describe symptoms that the patient suffers from.

3.2 Adding emotional responses and summaries

When a definition is added to the dialogue, a definition phrase is placed in a template that matches its syntactic category and a response is constructed that accords with its polarity. In a previous version of the system, medical information was presented through sequences of question-answer pairs, that is, questions about entities or relations from the simulator database and answers giving definitions. The new strategy presents the information as question-answer-response triples. These have the effect of slowing down the rate of communication of information, as suggested by our pilot experiment reported in [18]. The NLG system adds a definition question such as "Anaemia?", or "What is anaemia", or "What is that?", along with an answer that includes the negative polarity dictionary definition "anaemia is a condition in which patients feel very tired and may become breathless". Then it adds the emotional response: a neutral response for a neutral or negative polarity definitions is randomly chosen from "right". "okay", and "I see". A positive response for a positive polarity definition is randomly selected from "cool!", "amazing!" "I never knew!", and "just imagine it!". These come with Loquendo text-to-speech software as pre-recorded phrases; we chose these particular ones because their intonation accorded with the emotions that we wanted to convey, i.e., enthusiasm or concern.

The content planner also adds summaries of each medical episode (intervention or investigation) in the patient's record. These clarify and repeat the information. Summaries are of two kinds:

- **Authoratative.** Senior nurse summaries, e.g., "So, a packed red cell transfusion was administered to treat the anaemia."
- **Emotional.** Junior nurse summaries, e.g., "So, let's hope that the packed red cell transfusion took care of the anaemia."

Each embodied agent has a number of built-in gestures that can be associated with textual utterances so that a gesture will play at roughly the same time as a phrase is spoken. However, with Cantoche agents, synchronisation of speech and gestures cannot be fine-tuned to the extent where a gesture can be played to emphasise an individual word or syllable. Three types of gestures are generated by our current system: (a) generated randomly from a small set of fairly neutral speech gestures, e.g., a small raise of the hand, (b) nods or shakes of the head to accompany "yes" or "no" utterances, and (c) the junior nurse takes out a clipboard and reads from it when the senior nurse asks him a question about the patient's medical record.

4 RELATED WORK

The automated generation of dialogue scripts was pioneered by Elisabeth André and collaborators [1]. Extending this work, in the NECA project, script generation was brought together with multimodal NLG [10] and emotive speech synthesis [16] resulting in Fully Generated Scripted Dialogue (FGSD) [17]. The NECA system has a number of important similarities and differences with the current system. First of all, although the NECA platform was domain-independent, the domains to which it was applied, car sales (eShowroom) and social chat (Socialite), put demands on information presentation quite different from those in the medical domain.

Let us illustrate how evaluative comments are dealt with in NECA, following the approach explored first in [1], using the car sales domain. In the domain model, the values of attributes of cars (e.g., horse power, top speed) are given a valence (positive or negative) for each of the dimensions that a potential car buyer might be interested in

(i.e., sportiness, family friendliness, etc.). The system generates dialogues between a virtual car seller and buyer. They might discuss a particular attribute of a car that the user is interested in. Depending on the valence of the attribute and the attribute value, the system can generate evaluative remarks by the buyer character depending on the dimension that interest her (these can be selected by the user). For example, a seller and buyer might discuss the top speed of a particular car with the buyer asking for the top speed and the seller answering 'It has a top speed of 180 mph'. Depending on whether the buyer is interested in environmental friendliness or sportiness of cars, she might then respond with either, for instance, 'Interesting, but isn't that bad for the environment?' or 'Great, that's a very fast car!'.

A difference with the current medical scenario is that whereas in the NECA domains positive/negative valence translates directly to a positive/negative comment (though it is modulated by the personality of the character), in our junior/senior nurse dialogues there is an asymmetry between positive and negative polarity definitions: whereas definitions with a positive polarity attract a positive response, definitions with a negative polarity lead to a neutral response. The rationale is that with the viewer being the patient, emphasing negative information is emotionally insensitive: the aim is to avoid upsetting the viewer and to show sympathy and a positive attitude (enthusiasm) whereever possible and avoid negative emotions.

A further difference is that the ability of the NECA system to generate evaluative remarks was never evaluated; in particular, its relation to naturalness and engagement were not empirically tested. The nearest evaluation of affective natural language in NECA concerned a comparison of two referring expression generation strategies, one for egocentric and one for neutral speakers (see [12]).

More closely related to the current medical domain, the Text-to-Dialogue (T2D) system ([11]) generates dialogue scripts for two computer-animated characters – a pharmacist and a client. T2D, however, generates the scripts from textual input (Patient Information Leaflets) rather than data. Both approaches build on the idea put forward in [13] that (rhetorical) relations between spans of text or data often lend themselves for presentation through dialogue patterns – for example, a causal relation between informational items A and B can be expressed in a dialogue between layman L and expert E of the form $L : Why\ A?\ E : Because\ B$.

In recent years, the topic of affective NLG, in particular for embodied agents, has attracted a lot of interest (see [9] for an overview of work up to 2003; and [2, 15] for collections of papers on embodied agents including a number on generation of affective language). One of the early embodied agents for medical applications, Greta, is described in [8]. Greta is an embodied conversational agent that can play the role of doctor in information-delivering dialogues with patients. It integrates BDI (belief, desire and intention) planning with affective state generation and recognition, and makes use of sophisticated integrated realization of language and gestures that is sensitive to the emotions of the patient. The main difference with our approach is that it aims at direct interaction with the user through dialogue, rather than the use of dialogue between two virtual characters as a means for information delivery. Whereas the Greta agent takes into account whether it is speaking with a patient or a doctor (adjusting its display of emotions accordingly), it does not factor in the possibility of an overhearer who might listen in on a conversation between two doctors, and thus influence their use of language.

The 'Carmen's Bright Ideas' system ([7]) occupies the middle ground between interactive systems, such as Greta, and our system which is aimed purely at *presenting* dramatic dialogue. Carmen's Bright Ideas is intended for parents of children with cancer. It interactively generates dialogues between animated characters using pre-recorded speech. User have some control through clicking on alternative emotional "thought balloons", though the overall storyline is maintained by a director module. This system was subject to a trial in which it replaced a research assistant who was teaching Bright Ideas (a self-help philosophy) to sixteen learners in some of their sessions. Learners responded positively to questions about the helpfulness and clarity of the system.

5 EXPERIMENT

5.1 Materials

We generated a medical drama script from one patient's (simulated) data. The script – see the appendix for the complete script – contained the kinds of emotional reactions to medical information described above. We manually cut out some of the script so that it lasted approximately three minutes (in practice, we cut out repetitions of medical investigations and interventions, e.g., a cancer patient who undergoes chemotherapy often becomes anaemic and consequently has many blood tests and blood transfusions to correct this condition; in such cases we only kept the first occurrence of each type of investigation and intervention). We then recorded a video of the embodied agents "acting" the drama which was shown to participants in the "emotional reactions" group.

A second script was made by manually editing the first one. All emotional reactions to medical information were cut out and emotional summaries made by the junior nurse were replaced with neutral ones, e.g., "So, let's hope that the packed red cell transfusion took care of the anaemia" was replaced with the unemotional "So, a packed red cell transfusion was administered to treat the anaemia". Another video was recorded as before and it was shown to participants in the "no reactions" group.

We designed an on-line questionnaire to elicit judgements about nine statements with an on-line survey tool (www.surveymonkey.com). The statements were arranged into three groups, each on a separate Web page, and a final page where participants could type comments, as follows:

Page 1: The video captured my attention.

Page 2: The woman behaved naturally.
The woman sympathised with the patient.
The woman was cold and unfeeling.
The woman was enthusiastic about medical facts.

Page 3: The man behaved naturally.
The man sympathised with the patient.
The man was cold and unfeeling.
The man was enthusiastic about medical facts.

Page 4: Free text comments.

A set of judgements was associated with each statement ("Strongly disagree", "Disagree", "Disagree a bit", "Don't know", "Agree a bit", "Agree" and "Strongly agree") from which participants were able to select only one. Each judgement was associated with a numerical value on a Likert scale ranging from 1 = "Strongly disagree" to 7 = "Strongly agree".

5.2 Participants

Forty adults, thirty-two females and seven males, who are known by the first author, were invited to participate. They were randomly allocated to one of the two groups, "emotional reactions" or "no reactions", and were sent an e-mail asking them to participate and directing them to a Web page containing the materials relating to their group's condition. Thirty people completed the questionnaire.

5.3 Method

The participants watched a video on a Web site; they were able to view it as many times as they liked. Following successful viewing, they were redirected to another Web site where they were invited to respond to each of the above statements by selecting one judgement. The on-line questionnaire was set up so that participants could not proceed unless they selected a judement for each statement. Their selections were recorded as numerical values on a Likert scale as above. Responses to the questionnaire were collected anonymously by the on-line survey tool (www.surveymonkey.com). The tool records I.P. addresses and does not allow submission of more than one questionnaire from an I.P. address. Since the participants were known to us and because most of them also sent personal e-mails to let us know that they had completed the questionnaire, we are confident that the twenty-eight responses that we received are genuine and valid.

5.4 Results

The main issue is whether the inclusion of emotional reactions influenced viewers' judgements about (a) their interest in the video and (b) the attitudes and behaviour of the embodied characters. Table 1 shows mean judgements for each statement by the two groups (emotional reaction present/absent). As can be seen, the groups gave similar positive judgements on whether the video held their attention (5.13 vs 5.43, n.s.). However, significant differences (independent samples t-test) were found for two judgements (starred): when the man (the junior nurse) gave emotional reactions he was perceived as being more sympathetic towards the patient (4.88 vs 3.57, $p < 0.015$) and more enthusiastic about medical facts (5.06 vs 3.50, $p < 0.003$). Since the woman (the senior nurse) uttered very few emotional responses (apart from agreeing occasionaly with the junior nurse's hope that the treatment worked), we did not expect significant differences between the two conditions in perception of her attitudes.

Table 1. Mean judgements ranging over values from 1 (strongly disagree) to 7 (strongly agree).

Statement	Emotional Reaction n=16	No Reaction n=14
Video captured my attention	5.13	5.43
Woman behaved naturally	4.19	5.14
Woman sympathised with patient	4.31	4.00
Woman cold and unfeeling	2.94	2.71
Woman enthusiastic about medical facts	5.44	4.93
Man behaved naturally	4.06	3.57
Man sympathised with patient*	4.88	3.57
Man cold and unfeeling	2.94	2.93
Man enthusiastic about medical facts*	5.06	3.50

Table 2. Frequencies for Agree, Disagree, Don't know (n=30)

Statement	Agree	Disagree	Don't know
Video captured my attention*	25 (83%)	5 (17%)	0
Woman behaved naturally	20 (67%)	10 (33%)	0
Woman sympathised with patient	10 (33%)	9 (30%)	11 (37%)
Woman cold and unfeeling*	6 (20%)	22 (73%)	2 (7%)
Woman enthusiastic about * medical facts*	24 (80%)	4 (13%)	2 (7%)
Man behaved naturally	12 (40%)	17 (57%)	1 (3%)
Man sympathised with patient	11 (37%)	9 (30%)	10 (33%)
Man cold and unfeeling*	4 (13%)	22 (73%)	4 (13%)
Man enthusiastic about medical facts	18 (60%)	10 (33%)	2 (7%)

The results also show some tendencies that were common to the two groups. Table 2 gives frequencies for positive, negative and neutral responses to the statements, with data pooled so that each row sums to the total number of subjects (30). A judgement is classified as positive (Agree) if it lies in the range 5-7, negative (Disagree) if it lies in the range 1-3, and neutral (Don't know) if it is equal to 4. Overall there is a slight bias (130 vs. 108) for positive responses over negative; taking this into account, an agree-disagree split of 20:10 (or 10:20) has a probability $p < 0.02$ (binomial test) and a 25:5 split a probability of $p < 0.0004$ (binomial test), the starred comparisons are therefore significant. Inspection of the table reveals the following:

- Overall, the video succeeded in holding the viewers' attention, with responses largely positive.
- The characters were not seen as cold and unfeeling. Both for the woman (senior nurse) and the man (junior nurse), this statement was rejected with a significant split.
- The characters were seen as enthusiastic about medical facts, although this tendency was significant only for the woman. This is unsurprising since it was the woman who explained the medical terms. The perceived enthusiasm of the male was dependent on his emotional responses (see Table 1).
- Viewers were divided over whether the characters behaved naturally, with no significant differences, although neutral responses were rare (only one response in the 'Don't know' column).
- Viewers found it hard to make a judgement over whether the characters were sympathetic towards the patient. Overall, only 32 of 270 responses were 'Don't know', and the probability (binomial test) of obtaining as many as 11/30 such responses is significantly low ($p < 0.05$).

Still with data pooled across the two groups, table 3 shows correlations among the subjects' responses to the statements. Here, we think the point of major interest is the first column showing which judgements about the characters are most strongly related to judgements about whether the video was attention-worthy. The results suggest that the video held a subject's attention more when he/she thought the characters showed sympathy towards the patient, were not cold and unfeeling, and were natural (woman only); these correlations are significant ($p < 0.05$, Pearson two-tailed test).

Finally, free text comments were provided by nine participants. The content of these provided valuable clues to their perception of the agents' behaviour. The persistent questions of the male nurse about the meaning of medical terms motivated three people to note that he appeared remarkably ignorant and for one to comment that he seemed to have a poor grasp of English, or worse, poor comprehension which could be dangerous. One respondent thought the

Table 3. Pearson Correlations, n=30, 2-tailed significance in parentheses, attn = the video captured my attention, w = the female embodied agent, m = the male embodied agent, cold = the agent was cold and unfeeling, enth = the agent was enthusiastic about medical facts, nat = the agent behaved naturally, symp = the agent sympathised with the patient.

	attn	w_cold	w_enth	w_nat	w_symp	m_cold	m_enth	m_nat	m_symp
attn	-	-	-	-	-	-	-	-	-
w_cold	-.452*(.012)	-	-	-	-	-	-	-	-
w_enth	-	-.529**(.003)	-	-	-	-	-	-	-
w_nat	.430*(.018)	-.590**(.001)	.510**(.004)	-	-	-	-	-	-
w_symp	.368*(.046)	-.563**(.001)	.557**(.001)	-	-	-	-	-	-
m_cold	-.434*(.017)	.606**(.000)	-	-.516**(.004)	-	-	-	-	-
m_enth	-	-	-	-	-		-	-	-
m_nat	-	-	-	-	-	-	.483**(.007)	-	-
m_symp	.416*(.022)	-.487**(.006)	.414*(.023)	-	-	-.481**(.007)	.764**(.000)	-	-

wording of some of the male nurse's questions made him sound particularly stupid and suggested alternatives, some of which are already part of our system's set – clearly, rather than selecting the form of questions randomly, in future we should derive a better method for choosing appropriate formulations to suit different dialogue situations. Two people liked the female nurse's explicit definitions of technical terms, but whilst one of them liked the repetitions of definitions, the other thought that these should not be repeated verbatim, but reformulated (this is another good candidate for further investigation, but currently it is beyond the scope of our system). Regarding the video interface, one person liked being able to read the text of the speech from the Cantoche agents' speech bubbles, another disliked the background scene showing a desk and plant and two people had problems with slow download and synchronisation of speech and video – these sometimes occur with poor Internet access and different browser versions.

6 CONCLUSIONS

The fundamental purpose of the video is to instruct — to help patients pick up facts and terminology relevant to their condition. At the same time we obviously aim to avoid boring the patient, or giving offence. We have explored in this study the hypothesis that an instructive video will hold the viewer's attention better if the characters display sympathy for the patient and enthusiasm for the medical information given. The outcome does not directly support this hypothesis. By including emotional reactions by the junior nurse, we obtain a significant increase in subjects' ratings of his enthusiasm and sympathy, but no increase in the rating given to the video (i.e., the judgement on whether it held the attention).

Paradoxically, the correlation data (pooling the groups) seem to tell a different story. Here we find a clear indication that subjects who gave higher ratings for sympathy also gave higher ratings to the video. A possible resolution is that the emotional reactions had some effect in increasing attention to the video, but not large enough to override other influences that might vary considerably across small groups of subjects. In this connection, it is also important to note that in our study we had only a single item for each condition. As pointed out by [5], this calls into question any conclusions one might want to draw regarding the influence of the two conditions, because there is no control for random variations in the material that might have influenced the answers of the participants.

Another curious outcome is that sympathy for the patient, the character trait most influenced by the independent variable (presence/absence of emotional reaction), was also the trait that subjects found hardest to assess: out of a total of 60 responses to the sympathy questions, 21 fell into the 'Don't know' category, which was used only 11 times for all the other responses. It seems that subjects are strongly influenced by whether the characters show sympathy towards the patient, but found this hard to judge from the evidence of the video. Perhaps this was because we deliberately avoided any direct expressions of sympathy, for fear that subtle mistakes in tone might give offence. The lesson from our data is that this problem needs to be addressed, tricky though it is, since appropriate displays of sympathy would increase the viewer's attention to the video and its message.

As a final qualification, we should point out that the subjects in this experiment were not cancer patients. They were therefore judging the video, and its characters, in the role of outsiders with (perhaps) some general interest in medicine, rather than people personally affected by the material. However, our results should generalise to instructive videos ('edutainment') for use in education and training, even though special testing would obviously be needed before presentations of this kind could be used as a resource in treatment.

ACKNOWLEDGEMENTS

We would like to thank the two anonymous reviewers for their comments on a draft of this paper. This research was supported by Medical Research Council grant G0100852 under the e-Science GRID Initiative.

REFERENCES

[1] E. André, T. Rist, S. van Mulken, M. Klesen, and S. Baldes, 'The automated design of believable dialogues for animated presentation teams', in *Embodied Conversational Agents*, eds., Justine Cassell, Joseph Sullivan, Scott Prevost, and Elizabeth Churchill, 220–255, MIT Press, Cambridge, Massachusetts, (2000).

[2] J. Cassell, J. Sullivan, S. Prevost, and E. Churchill, *Embodied Conversational Agents*, MIT Press, Cambridge, Massachusetts, 2000.

[3] R. Cox, J. McKendree, R. Tobin, J. Lee, and T. Mayes, 'Vicarious learning from dialogue and discourse: a controlled comparison', *Instructional Science*, **27**, 431–458, (1999).

[4] S D. Craig, B. Gholson, M. Ventura, and A Graesser, 'Overhearing dialogues and monologues in virtual tutoring sessions: Effects on questioning and vicarious learning', *International Journal of Artificial Intelligence in Education*, **11**, 242–225, (2000).

[5] D. Dehn and S. van Mulken, 'The impact of animated interface agents: a review of empirical research', *Int. J. Human-Computer Studies*, **52**, 1–22, (2000).

[6] A. King, 'Guiding knowledge construction in the classroom: Effects of teaching children how to question and how to explain', *American Educational Research Journal*, **31**, 338–368, (1994).
[7] S. Marsella, W.L. Johnson, and C. LaBore, 'Interactive pedagogical drama for health interventions', in *11th International Conference on Artificial Intelligence in Education, AIED 2003*, (2003).
[8] C. Pelachaud, V. Carofiglio, B. De Carolis, F. de Rosis, and I. Poggi, 'Embodied Contextual Agent in Information Delivering Application', in *Proceedings of the First International Joint Conference on Autonomous Agents and Multi-Agent Systems*, Bologna, (2002).
[9] P. Piwek, 'An annotated bibliography of affective natural language generation', ITRI Technical Report ITRI-02-02, ITRI, University of Brighton, (2002). Version 3 (2003) Available at http://mcs.open.ac.uk/pp2464/affect-bib.pdf.
[10] P. Piwek, 'A flexible pragmatics-driven language generator for animated agents', in *Proceedings of 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL) (Research Notes)*, pp. 151–154, Budapest, Hungary, (2003).
[11] P. Piwek, H. Hernault, H. Prendinger, and M. Ishizuka, 'T2D: Generating Dialogues between Virtual Agents Automatically from Text', in *Intelligent Virtual Agents: Proceedings of IVA07*, LNAI 4722, pp. 161–174. Springer Verlag, (2007).
[12] P. Piwek, J. Masthoff, and M. Bergenstrahle, 'Reference and Gestures in Dialogue Generation: Three Studies with Embodied Conversational Agents', in *Proceedings of AISB05 Virtual Social Agents Symposium*, University of Herfordshire, (2005).
[13] P. Piwek, R. Power, D. Scott, and K. van Deemter, 'Generating Multimedia Presentations from Plain Text to Screen Play', in *Multimodal Intelligent Information Presentation*, volume 27 of *Text, Speech and Language Technology*, 203–225, Springer, Dordrecht, (2005).
[14] J. Preece, Y. Rogers, and H. Sharp, *Interaction Design: Beyond Human-Computer Interaction*, John Wiley & Sons, New York, NY, 2002.
[15] H. Prendinger and M. Ishizuka, *Life-Like Characters: Tools, Affective Functions, and Applications*, Cognitive Technologies Series, Springer, Berlin, 2004.
[16] Marc Schröder, *Speech and emotion research: an overview of research frameworks and a dimensional approach to emotional speech synthesis*, Ph.D. dissertation, Institute of Phonetics, University Saarland, 2004.
[17] Kees van Deemter, Brigitte Krenn, Paul Piwek, Martin Klesen, Marc Schröder, and Stefan Baumann, 'Fully generated scripted dialogue for embodied agents', *Artificial Intelligence Journal*, **to appear**, (2008).
[18] Sandra Williams, Paul Piwek, and Richard Power, 'Generating monologue and dialogue to present personalised medical information to patients.', in *11th European Workshop on Natural Language Generation*, pp. 167–170, (2007).
[19] Claus Zinn, Johanna D. Moore, and Mark G. Core, *Intelligent Information Presentation*, chapter Intelligent Information Presentation for Tutoring Systems, Kluwer, 2005.

7 APPENDIX

Generated dialogue

The dialogue below was used in the experiment. Emotional responses generated by the system are marked (for the condition with no emotional responses, either there was no response, or alternative unemotional responses were generated as shown).

Senior: Could you read the patient's notes for June 21st.
Junior: She had an excision on June 21st.
Junior: excision?
Senior: An excision is the removal of the cancer along with a border of healthy tissue around it.
Junior: Would you repeat it?
Senior: An excision is the removal of the cancer along with a border of healthy tissue around it.

Junior (emotional response): right

Senior: What was treated?
Junior: A tumour.
Senior: Where was the tumour?
Junior: The tumour was in her left breast.
Senior: So, an excision was administered to treat the cancer.

EITHER
Junior (emotional response): And we hope it worked.
Senior: Yes.
OR
Junior: Uh-huh

Senior: Go on.
Junior: From September 27th to October 25th she had a radiotherapy course.
Junior: radiotherapy course?
Senior: Radiotherapy targets cells with radiation.

Junior (emotional response): Cool!

Senior: So, a radiotherapy course was administered to treat the cancer.

EITHER
Junior (emotional response): And we hope it worked.
Senior: ok.
OR
Junior: Uh-huh

Senior: And then?
Junior: Her record says that she had a packed red cell transfusion.
Junior: packed red cell transfusion?
Senior: A packed red cell transfusion is a transfusion of red blood cells.
Senior: Are you following me?
Junior: ok
Senior: Red blood cells contain haemoglobin which carries oxygen around the body.

Junior (emotional response): I never knew!

Senior: And that treatment was for?
Junior: Anaemia.
Junior: anaemia?
Senior: Anaemia is a condition in which patients feel very tired and may become breathless.

EITHER
Junior (emotional response): okay
Junior (emotional summary): So, let's hope that the packed red cell transfusion took care of the anaemia.
OR
Junior: So, the packed red cell transfusion treated anaemia.

Senior: Uh-huh
Senior: Continue please.
Junior: Her record says that she had an examination.
Senior: Where?
Junior: The axillary lymphnodes.
Junior: What are axillary lymphnodes?
Senior: The axillary lymphnodes are the rounded masses of tissue under the arms containing white blood cells.

Junior (emotional response): just imagine it!

Senior: What did the examination reveal?
Junior: lympha, lympha...
Senior: Does it say lymphadenopathy?
Junior: Yes
Junior: What is that?
Senior: Lymphadenopathy is a swelling of the lymph nodes which the doctor can feel when you are examined.
Junior: What did you say?
Senior: Lymphadenopathy is a swelling of the lymph nodes which the doctor can feel when you are examined.

Junior (emotional response): I see

Senior: So, an examination led to detection of the lymphadenopathy.
Junior: Uh-huh.