ANTHROPOMORPHIC VS NON-ANTHROPOMORPHIC USER INTERFACE FEEDBACK FOR ONLINE HOTEL BOOKINGS

Pietro Murano, Anthony Gee
Computing Science and Engineering, University of Salford, Newton Building, Gt. Manchester, M5 4WT
p.murano@salford.ac.uk, tonygee47@hotmail.com

Patrik O'Brian Holt
Interactive Systems Research Group, School of Computing, The Robert Gordon University, St. Andrew Street, Aberdeen, AB25 1HG, Scotland
p.holt@rgu.ac.uk

Keywords: Anthropomorphism, User interface feedback, Evaluation.

Abstract: This paper describes an experiment and its results concerning research that has been going on for a number of years in the area of anthropomorphic user interface feedback. The main aims of the research have been to examine the effectiveness and user satisfaction of anthropomorphic feedback in various domains. The results are of use to all interactive systems designers, particularly when dealing with issues of user interface feedback design. There is currently some disagreement amongst computer scientists concerning the suitability of such types of feedback. This research is working to resolve this disagreement and in turn can help software houses to increase their profits by developing better user interfaces that will promote an increase in sales. The experiment detailed, concerns the specific software domain of Online Factual Delivery in the specific context of online hotel bookings. Anthropomorphic feedback was compared against an equivalent non-anthropomorphic feedback. Statistically significant results were obtained suggesting that the non-anthropomorphic feedback was more effective. The results for user satisfaction were however less clear.

1 INTRODUCTION

User interfaces and the feedback given to users are one of the most important aspects of any software system. This is because if the user interface and the feedback given is not usable, the users will either give up using the system, will be less efficient in using the system or will simply not enjoy using the system. This in turn can seriously affect the success of a software house and its sales. Also the growth and complexity of modern day software systems, in particular the tasks they are able to perform, results in the continual requirement for more usable interfaces to be developed.

The aim of this research is to aid in the improvement of user interfaces for users which can promote better sales for a software house. Specific concentration is placed on comparing anthropomorphic and non-anthropomorphic user interfaces to address the issues of effectiveness and user satisfaction.

There are various opinions amongst the computer science community regarding the effectiveness and user approval of anthropomorphic feedback at the

user interface. Some researchers are in favour of anthropomorphism, e.g. Koda and Maes (1996), Maes (1994), Laurel (1997), Agarwal (1999), Zue (1999) and Takeuchi and Naito (1995) However, some researchers are not generally in favour of anthropomorphism in most circumstances e.g. Shneiderman and Plaisant (2005). Each of these researchers tends to base their opinions on various studies conducted in the area. Due to the inconclusive nature of the results of these studies, there is the need for more work in this area to gain a better understanding.

This research continues on from a number of research studies conducted by Murano (2005, 2003, 2002a, 2002b, 2001a, 2001b) aiming to eventually solve the issues of effectiveness and user satisfaction of anthropomorphic feedback at the user interface. In Murano (2002b) it was shown that in the domain of software for in-depth learning, anthropomorphic feedback was significantly more effective. The results for user satisfaction were not so clear, but participant preferences tended towards the anthropomorphic feedback. This was specifically in the context of English as a Foreign Language pronunciation. Also in Murano (2002a) it was shown that in the domain of software for online systems usage, anthropomorphic feedback was significantly more effective and preferred by users. This context specifically involved the area of using UNIX commands at the UNIX shell.

Specifically related to this paper, are the results in Murano (2003). The paper investigated anthropomorphic feedback in the context of online factual delivery, using the area of direction finding as the specific context. This paper showed with statistically significant results, that non-anthropomorphic feedback was more effective. The results for user satisfaction were not so clear, but participant preferences tended towards the non-anthropomorphic feedback.

In Dehn and van Mulken (2000) it was suggested that the context or domain of concern could influence the effectiveness and user approval of anthropomorphic interfaces. This research is beginning to suggest with empirical evidence that this could be the case. However, in order to make sure that this really is the case, the authors are investigating anthropomorphic feedback in different domains, with the possibility of eventually developing a taxonomy of feedbacks as suggested in (Murano, 2005), for helping user interface designers in their design decisions.

This paper therefore investigates the domain of online factual delivery further, describing an experiment set in this domain, using the context of online hotel bookings to test the user interface feedback. This context was chosen because it is a fairly common activity for users of all kinds to carry out over the Internet and was therefore considered to be useful and realistic, whilst maintaining the theme of the previous experiment conducted by Murano (2003). As with the previous experiments, effectiveness and user satisfaction were the aspects being investigated. Effectiveness was defined by the success rate in completing the tasks, a low error rate whilst carrying out the tasks and a low rate of hesitations/frustrations expressed by the participants during the experiment. The user approval aspects concerned the participants' subjective opinions regarding the user interface aspects. For this experiment, the anthropomorphic feedback consisted of an animated character supplied with MS Agent 2.0 (see Apparatus and Material section) called 'Merlin'. The non-anthropomorphic feedback consisted of guiding text. This was text of the kind one would expect to see on a 'real' online hotel booking site.

2 THE EXPERIMENT – HOTEL BOOKINGS

2.1 Hypotheses

As stated in the previous section this research concerns determining the effectiveness and user satisfaction of anthropomorphic user interface feedback in various contexts. Hence the following hypotheses were derived:

H0a - There will be no difference in terms of user satisfaction between the anthropomorphic feedback (Merlin) and non-anthropomorphic feedback (guiding text).

H0b - There will be no difference in terms of effectiveness between the anthropomorphic feedback and non-anthropomorphic feedback.

Positive Hypotheses:

H1a - The non-anthropomorphic (guiding text) feedback will be more effective than the anthropomorphic (Merlin) feedback.

H1b - Users will prefer the anthropomorphic (Merlin) feedback rather than the non-anthropomorphic (guiding text) feedback.

2.2 Pilot Testing

Before the main experiment was undertaken, a small pilot test with 4 participants was conducted. The main issues being considered in the pilot test were the main workings of the prototype developed, the environment to be used in the experiment and exercising suitable control over the various variables being tested (see Variables section). A further aspect aided by the pilot test, was to determine a suitable amount of time to be used for the experiment and to test out the actual designed tasks.

2.3 Users

The initial recruitment of the participants took place by means of a recruitment questionnaire. The participants were carefully selected so as to have similar profiles, therefore reducing the possibility of collecting invalid data. Initially 40 individuals were selected, but only 20, with similar profiles, were actually used in the experiment. The main aspects of the profiles of the participants used were similar in the following ways:

- All participants had similar computing knowledge. They were not complete beginners or 'power' users. Complete novice users were not selected as they would have required basic training in the concepts of devices and Windows systems. Experienced participants were not used in the experiment as it was decided that such users would in reality not require feedback of the sort being tested in their every day usage patterns.
- All the participants were less than 36 years of age with English as their primary language.

2.4 Experimental and Task Design

For the purpose of the given experiment a between users design method was deployed. 10 of the participants were assigned to Group A, and the remaining 10 participants were assigned to Group B.

Group A participants tested the anthropomorphic feedback (MS Merlin) as part of their experiment session.

Group B participants tested the non-anthropomorphic feedback (guiding text) as part of their experiment session.

The experiment involved each participant attempting the following tasks:

- Task 1 required participants to make a specific booking for a hotel and theatre performance. Participants would use the prototype online hotel reservation user interface to make the bookings according to specific details supplied.
- Task 2 required participants to cancel the booking they had just made using the hotel reservation user interface.

The tasks outlined are representative of realistic tasks commonly carried out by users booking a hotel or holiday, using the Internet. For tasks 1 and 2 all participants were initially shown a brief tutorial explaining how to book and cancel a hotel using the interface. The content of the tutorials shown was identical regardless of the feedback being given to ensure there was no bias.

2.5 Variables

For the purpose of the experiment the associated independent variables were determined as being the two different methods of feedback that were available:

- Animated Microsoft Merlin with speech and text (anthropomorphic).
- Standard guiding text (non-anthropomorphic).

The dependent variables were the participants' performance in dealing with the hotel bookings and their subjective opinions.

The dependent measures were that the performance was measured by counting the number of errors incurred, observing whether participants completed the tasks and counting the number of times hesitation and frustration were manifested. These factors were then used in a scoring formula (see Scoring section below for a description of the formula). Specifically performance was measured in the following manner:

- Tasks carried out correctly with no errors. The participants were given a task sheet with specific instructions regarding the booking they should make (e.g. given dates and number of guests etc.). Deviation from this was considered to be a complete task but with some incorrect details.
- Tasks completed. This refers to the overall successful completion of the two tasks in the experiment.
- Number of times participants showed signs of hesitation. These were only clearly observable participant reactions, such as manifesting a puzzled expression or asking for help.
- Number of times participants showed signs of frustration. These were only clearly observable user reactions, such as making some remark about the user interface which had clearly caused the user some 'anger'.
- The number of times participants used the feedback help, but still made an error.

These factors were recorded by means of an observation protocol.

The subjective opinions were measured by means of a post-experiment questionnaire. Participants were asked to rate various aspects of the user interface using a Likert scale, where 9 was the most positive score regarding some opinion, and 1 was the most negative score available.

2.6 Apparatus and Materials

The experiment involved the use of 'standard' equipment. These were a laptop with, 128MB RAM, 20Gb disk and Windows XP. Also Microsoft Agent 2.0, the "Merlin" character and Lernout & Hauspie TruVoice Text-to-Speech (TTS) engine (American English) were used. Supplementary hardware used consisted of an external mouse and external speakers. Further, a paper notepad was available for each participant, for use in the experiment (see Procedure section). Each prototype was developed using Visual Basic 6. The Anthropomorphic interface required the use of the Microsoft Agent 2.0 Active X component.

2.7 Procedure

The first process was to recruit suitable experiment candidates. This involved utilising the participant recruitment questionnaire to ask specific questions regarding the participants' background and experiences, to determine whether the participant met the selection criteria. Once all the suitable participants were recruited, they were randomly assigned into either Group A or Group B (see Experimental and Task Design section). Participants were then contacted and asked to meet at a suitable time to take part in the experiment.

The experiment itself took approximately 30 minutes to complete. The procedure involved ensuring that each participant was treated in the same way, with the following outlined procedure being identical for each of the participants. Also all the questionnaires and observation techniques used were the same for each participant, with the aim of minimizing confounding variables.

The experiment took place in a carefully controlled environment, ensuring that there were no distractions and that the participants felt at ease.

Upon entering the room each participant was greeted by the experimenter and was made to feel comfortable and relaxed. To make them feel more at ease, light refreshments were also offered at this time. The participants received a short verbal introduction to the experiment, explaining the purpose of the study, with reassurance that the software was the focus of the study and not themselves. At this time participants were informed that they would be observed by the experimenter who would be present in the room throughout the experiment. When the participant felt relaxed, a task sheet was given to them, which contained a brief introduction to the experiment along with Tasks 1 and 2 (see Experimental and Task Design section). Having read through the task sheet participants were again assured that they were not being examined and they were subsequently asked if they had any immediate concerns regarding the tasks. Participants were then instructed as to which method of feedback they would be testing.

Once the participant was ready the program began with a brief tutorial using the relevant method of feedback (Group A anthropomorphic and Group B non-anthropomorphic in terms of feedback). Both tutorials, regardless of the feedback, were the same in content. The only differences involved the anthropomorphic character referring to itself as 'I', while the non-anthropomorphic feedback was neutral in nature. The tutorial informed the participant how to book and cancel a hotel using the prototype. When the tutorial was started, the relevant mode of feedback 'explained' how to use each screen and its features. All the screens involved in the tutorial dealt with bookings and the cancellation of bookings. For the anthropomorphic condition the character uttered the information and this was also concurrently viewable by means of corresponding speech bubbles. Further, the character moved on the screen and 'pointed' with a hand to the features of each screen as it was being 'described'. For the non-anthropomorphic condition, the same information appeared in text boxes with arrows pointing to the various features of the screens.

Upon completion of the tutorial participants were then asked whether the tasks were clear, and when the participants felt ready the first task began.

Upon completion of task 1 participants were asked if they had any immediate comments as to the task they had completed, such comments being recorded in the observation notes. The participants were then asked whether they were ready to begin task 2, once comfortable, task 2 proceeded. It was determined that the task was complete when the participants had successfully cancelled the booking they had made during task 1. Following the task, completion comments and opinions were sought

from the participants.

Errors were categorised by recording whether a participant completed the task according to the specifications given on the task sheet. If the participants deviated from the instructions given, e.g. the hotel was booked for the party arriving on the wrong day, or not enough rooms booked etc, this was recorded as a participant completing a task but with some incorrect details (see Variables section above).

At times when the participants hesitated as to what they were required to do at a particular point, these hesitations were recorded (see Variables section above). At any point during the experiment if a participant asked the experimenter present in the room for guidance, no additional help was given. Instead participants were instructed that they should consult the feedback integrated into the interface, which was of the same kind as found in the tutorial and had the same condition being tested. If at any time a participant did consult the feedback, and still subsequently made an error regarding the problem they were trying to overcome, this was recorded. However, if the participant did consult the feedback and this solved the problem, this was also recorded. The number of times participants expressed clear frustration was also recorded. Such frustration included occurrences where participants would make remarks regarding certain aspects of the interface or feedback that caused them anger.

A particular aspect of the second task was to enter the booking reference supplied when participants made a booking during Task 1, so that the correct booking information could be retrieved to enable the booking to be cancelled. If a participant was unable to remember the booking reference (the software instructed the participant to note the reference during Task 1), having not written it down on the notepad provided, this would be seen as an error and subsequently resulted in the participant not fully completing Task 2.

Once all tasks had been completed the experimenter debriefed each participant. This included the completion of the post experiment questionnaire, elicitation of participants' immediate comments as well as the experimenter informing the participant how the results of the study will be made available if required.

2.8 Scoring

The effectiveness variables described (see Variables section) were carefully recorded for each participant. For each task completed/not completed, a score was assigned for use in the statistical analyses. The score for each task was based on a similar points system as published in Murano (2002a). For each task, each participant (unknown to them) was started on 10 points.

Events which caused the score to reduce were observations of the following types: Signs of frustration (negative physical attitude) or hesitation resulted in 0.5 points being deducted from the score. If the participant carried out an incorrect action, causing the system to display an error message, 0.5 points were deducted. If the participant consulted the feedback in a particular situation and despite the help, continued to make a mistake, 0.5 points were deducted from the running score.

Occurrences when the participant had completed the task but made a mistake in the booking, resulted in 1.5 points being deducted from the score. If the participant was unable to complete the task, 1.5 points were deducted. Finally if the participant completed the task with none of the noted penalties the score would remain at 10.

Consequently, at the end of each task the participant obtained a final score.

The formula was devised because it was felt that all the factors being measured potentially had a direct effect on overall success.

2.9 Results

The data obtained for this experiment concerned effectiveness and subjective user opinions issues. The effectiveness issues were statistically analysed using a t-test and the subjective opinions regarding the interface used, were analysed through their means and standard deviations.

For the 20 participants, 10 using the anthropomorphic feedback (MS Merlin) and 10 using the non-anthropomorphic feedback (guiding text), data gathered from the first task showed a t-observed of 3.08 and the t critical (5%) was 2.10, Table 1 below illustrates these statistics:

Table 1: T-test result of text Vs Merlin (task 1).

t-Observed	3.09
t-Critical (5%)	2.10

For the second task with 20 participants, 10 using anthropomorphic feedback (MS Merlin) and 10 using the non-anthropomorphic feedback (guiding

text), comparing between the two feedbacks the t-observed was 2.55 and the t critical (5%) was 2.10. Table 2 below illustrates these statistics:

Table 2: T-test result of text Vs Merlin (task 2).

t-Observed	2.55
t-Critical (5%)	2.10

For the 20 participants, 10 using anthropomorphic feedback (MS Merlin) and 10 using the non-anthropomorphic feedback (guiding text), across both tasks, the t-observed was 4.93 and t critical (5%) was 2.10. Table 3 below illustrates these statistics:

Table 3: T-test result of text Vs Merlin (tasks 1 and 2 combined).

t-Observed	4.93
t-Critical (5%)	2.10

Several subjective questions regarding the general user interface (e.g. buttons, screen sequencing and text clarity etc.) were asked of all participants. All subjective responses indicated that there were no negative issues or severe 'dislikes' with the general user interface.

The main aspects of importance regarding the participants' subjective opinions concerned the two types of feedback being tested, particularly in relation to the material presented in the initial tutorial and feedback given as part of the help sub-system. The relevant means and standard deviations can be seen in tables 4 and 5 below:

Table 4: Means and standard deviations (SD) for tutorial user subjective opinions.

Non - Anthropomorphic Group	**Mean**	**SD**
Tutorial Helpfulness	8.10	0.74
Detailed Tutorial	8.40	0.52
Tutorial Clarity	8.40	0.52
Satisfying Tutorial	7.10	0.74
Structure of Tutorial	7.80	0.79
Aided in Completing Task	7.80	0.79
Anthropomorphic Group	**Mean**	**SD**
Tutorial Helpfulness	7.50	1.08
Detailed Tutorial	7.70	0.48
Tutorial Clarity	8.40	0.52
Satisfying Tutorial	7.90	0.57
Structure of Tutorial	8.20	0.63
Aided in Completing Task	7.30	1.06

Table 5: Means and standard deviations (SD) for system help user subjective opinions.

Non-Anthropomorphic Group	**Mean**	**SD**
Usefulness of Help	8.20	0.63
Help Clarity	7.70	0.82
Stimulating Help	7.80	0.63
Adequate Help	8.10	0.57
Relevant Help	8.60	0.52
Quality of Help	7.80	0.79
Aided in Solving Problem	7.90	0.57
Help Understandability	8.40	0.52
Anthropomorphic Group	**Mean**	**SD**
Usefulness of Help	8.10	0.57
Help Clarity	8.00	0.47
Stimulating Help	8.00	0.67
Adequate Help	7.70	0.67
Relevant Help	8.20	0.79
Quality of Help	8.40	0.52
Aided in Solving Problem	8.00	0.47
Help Understandability	8.50	0.71

Participants were also asked their opinions regarding potential future use of the interface feedback they used during the experiment. For the non-anthropomorphic group (guiding text), 6 out of 10 participants said they would use the feedback again if made available. For the anthropomorphic group (MS Merlin), 8 out of 10 participants said they would use the feedback again if made available.

3 CONCLUSIONS

Initially tasks 1 and 2 were statistically analysed on an individual basis as task 2 was a shorter and easier task to carry out, so it was necessary to assess if this influenced the results in any way. The 2 tasks were also analysed together to give an overall assessment of the feedback.

The results from the individual tasks and the combination of the 2 tasks show clear statistical significance in favour of the non-anthropomorphic (guiding text) feedback. Participants completed the

tasks more successfully and with less errors/hesitations in the hotel booking context.

Consequently, with reference to the hypotheses stated earlier in this paper, it is now possible to reject the (H0b) null hypothesis, with the results showing that there is a clear difference between the feedbacks in terms of effectiveness. Statistical significance in the results enables the (H1a) positive hypothesis to be accepted. This postulated that the non-anthropomorphic feedback would be more effective.

Assessment of the user satisfaction of the two types of feedback in terms of the tutorial and help sub-system, shows that the 2 groups of participants each rated these fairly closely to each other, as can be seen by the means (tables 4 and 5). Also in all cases the scores provided by the participants are consistent, as the standard deviations are all rather low. Further there was not much difference in the responses of the 2 groups of participants, concerning whether the participants would be prepared to use the same feedback again, should it become available.

Therefore the null hypothesis (H0a) is accepted, the scores do not show enough difference between them to allow a different conclusion. The positive hypothesis (H1b) is therefore rejected.

These results follow the results obtained by Murano (2003) in the different context of direction finding – within the area of online factual delivery. The suggestion being made is that software for online factual delivery in some different contexts or domains is better suited to having a non-anthropomorphic type of user interface feedback.

Some interesting comments and observations were made by the participants during and after the experiment. The anthropomorphic feedback seemed to 'fascinate' some of the participants in this group. Some commented that they 'sat back' and 'watched'. Some also commented that they felt they were not actually learning anything. Certain individuals seemed to concentrate more on the 'appearance' of the Merlin character rather than concentrating on the words being uttered. These participant comments were reasonable as their observed behaviour matched their self-evaluation. Another interesting aspect concerns the fact that some participants in the anthropomorphic group stated that their experience with this feedback was 'engaging', 'involving' and 'fun'. The converse was true of some of the comments made by the non-anthropomorphic group, where some stated their experience was 'uninspiring' and 'normal'. These aspects were also evident as the participants were being observed. These participant comments and observations could explain why the anthropomorphic feedback was rated very closely to the non-anthropomorphic feedback. It could simply be that the anthropomorphic feedback had more of a novelty factor. However the authors suggest that this novelty factor would disappear with regular use of such a system.

These results are very important for user interface designers, as it would be the ideal scenario to be able to generalise these results to all types of software in the area of online factual delivery. Therefore, the experimental results are suggesting that for software for online factual delivery, non-anthropomorphic feedback is potentially more effective in terms of reducing user errors and hesitations. Users however rate both kinds of feedback highly, as was also found in Murano (2003). The suggestion could be therefore to have some element of anthropomorphic feedback along with the non-anthropomorphic feedback. This could be done by having the anthropomorphic feedback in a non-crucial role, while using the non-anthropomorphic feedback for the important aspects of an interaction. A further suggestion would be to make both kinds of feedback available to the user, by some 'toggle' function. Whichever strategy would be followed, it would require piloting in a real environment with potential real users. If an appropriate strategy could be found, the usability of an application could be enhanced bringing benefit to a software house involved in its development and to the user community.

ACKNOWLEDGEMENTS

The authors would like to thank The School of Computing, Science and Engineering at the University of Salford, Prof. Ritchings, Heriot-Watt University, Edinburgh, Dept. of Computer Science and The Robert Gordon University, Aberdeen, School of Computing are thanked for their support.

REFERENCES

Agarwal, A., 1999 Raw computation, *Scientific American*, 281, pp. 44-47.

Dehn, D. M., van Mulken, S., 2000 The Impact of animated interface agents: A Review of Empirical Research. *International Journal of Human-Computer Studies 52: 1-22.*

Koda, T., Maes, P., 1996 Agents with faces: the effect of personification, *Proc. of the 5th IEEE International Workshop on Robot and Human Communication (RO-MAN '96)*, pp. 189-194.

Laurel, B., 1997 Interface agents: metaphors with character. *In Software Agents*, ed Bradshaw, J.M. MIT Press, London.

Maes, P., 1994 Agents that reduce work and information overload, *Communications of the ACM*, 37(7), pp. 31-40.

Murano, P., 2005 Why anthropomorphic user interface feedback can be effective and preferred by users, *7th Int. Conference on Enterprise Information Systems. (c) – INSTICC*, Miami.

Murano, P., 2003 Anthropomorphic vs non-anthropomorphic software interface feedback for online factual delivery, *Proc. of the 7th. Int. Conference on Information Visualisation (IV'03)*, IEEE, p. 138, London.

Murano, P., 2002a Anthropomorphic vs non-anthropomorphic software interface feedback for online systems usage, *7th ERCIM Int. Workshop on User Interfaces for All*, pp. 339-349, Paris.

Murano, P., 2002b Effectiveness of mapping human-oriented information to feedback from a software interface *Proc. 24th International Conference on Information Technology Interfaces*, Cavtat - Croatia.

Murano, P., 2001a A new software agent 'learning' algorithm, *People in Control: An International Conference on Human Interfaces in Control Rooms, Cockpits and Command Centres*, IEE, Manchester.

Murano, P., 2001b Mapping human-oriented information to software agents for online systems usage, *People in Control: An International Conference on Human Interfaces in Control Rooms, Cockpits and Command Centres*, IEE, Manchester.

Shneiderman, B., Plaisant, C., 2005 *Designing the user interface: strategies for effective human computer interaction*, Pearson Education.

Takeuchi, A., Naito, T., 1995 Situated facial displays: towards social interaction, *Proc. CHI'95 Human Factors in Computing Systems*, pp. 450-454.

Zue, V., 1999 Talking with your computer, *Scientific American*, 281, pp. 40-41,