The 1^{st} Release of the TERENCE Learner GUI: the User-based Usability Evaluation

Cecilia M. R.[1], Di Mascio T.[2], and Melonio A.[3]

Abstract This paper reports the user-based usability evaluations performed in Italy of the first release of the learner Graphical User Interface (GUI) of the TERENCE project. This project aims at developing an adaptive learning system for training the reasoning about stories' events of the TERENCE learners in Italy and in UK. Learners are 7-11 year old children, hearing and deaf, that have difficulties in correlating the events of a story, making inferences about them, and detecting inconsistencies. The evaluation of the first release of the TERENCE adaptive learning system software prototypes tackles their usability in order to quickly reveal possible usability problems, as well as to address the TERENCE team to solve them, before the large scale evaluation. Moreover, authors try to carried out important general issues related to the experiment performance.

1 Introduction

The main reason to concentrate our effort on evaluating the usability of the TERENCE Graphical User Interfaces (GUIs) before the large scale evaluation mainly derives from the fact that, as well described in the [3] survey, "...the approaches used to evaluate Adaptive Learning Systems (ASLs) are similar in one aspect: they

[1] Cecilia Maria Rosita
Dep. of Life, Health and Environmental Sciences, University of L'Aquila, 67100, L'Aquila, Italy
e-mail: mariarosita.cecilia@univaq.it

[2] Di Mascio Tania
Dep. of Engineering, Information Science and Mathematics. University of L'Aquila, 67100, L'Aquila, Italy
e-mail: tania.dimascio@univaq.it

[3] Melonio Alessandra
Free University of Bozen-Bolzano, CS Faculty, P.zza Domenicani 3, 39100 Bolzano, Italy
e-mail: alessandra.melonio@inf.unibz.it

tend to evaluate an ALS as *a whole*, focusing on an *end value* delivered by the system such as the overall user's performance or the user's satisfaction... Evaluating a system as a whole can be acceptable in the field where no acceptable component model of a system can be identified. However, it is not the case for adaptive systems...'.

This paper suggests using a *layered evaluation process*, in which one of the layer is represented by the learning material and an other by GUIs. The authors introduced such an approach to guide designers in the ALS development process. Such a layered approach is in line with the *User-Centred Design* (UCD), used in the TERENCE project, where the evaluation is used iteratively and incrementally to refine the requirements, the design or the development of the system. Moreover, all the evaluation studies reported in [3] stressed the fact that the usability issues of the ALS interfaces have to be solved before starting the evaluation of the ALS in order to minimize bias in the evaluation study of the ALS' usability as "a whole", that is the ALS' pedagogical effectiveness. A thing which is mandatory in UCD as well.

The TERENCE project took up such a two layer up for the learning material and the GUIs, before the large scale evaluation, in two main manners:

- the TERENCE team evaluated the learner material and the GUI prototypes via expert-based evaluations reported in [9],
- the TERENCE team evaluated the refined version of the learner material and the first releases of the GUIs via user-based evaluation, reported in [5].

In this paper, we focus on the user-based evaluation, mainly reported in [5], of the most complex GUI, namely TERENCE learner GUI. The entire learner GUI is available at http://hixwg.univaq.it/learner-gui.html; its design is described in [10].

2 Experiment Description

For the experiments we here describe, we adopted user-based criteria methods like observational evaluation [1], semi-structured interviews [7] and think-aloud protocol [6]. In fact, the approaches used in the literature for evaluating TEL projects are mainly user-based (see [8]). An important reason justifying the usage of user-based approaches in TEL projects is the fact that users are often involved in the design of the projects. Like the other TEL projects, the TERENCE project involves users in the evaluation process. In doing so, the TERENCE team opted for methods that are adequate to the TERENCE main users, that is, 7–11 year old children, and prone to being used in numerous but short inexpensive evaluation sessions. In fact, observational evaluation, semi-structured interviews, and think-aloud protocol are semi-structured methods for examining and reporting problems with the learner GUI in qualitative and quantitative ways.

The reports of the assessments for the learner GUIs usability evaluations and the learning material in [5] is divided as follows: (1) *goals* of the assessment; (2) *participants*, that is, the description of the involved users; (3) *tasks and material*, that is,

the description of tasks and material proposed for the experiments; (4) *results*, that is, the description of significant results. This choice is the same we use to structure this section.

2.1 Experiment Goals

The overall goal was to examine whether the sequence of tasks in Table 1 and, more in general, the navigation of the learner GUI were usable for the intended age range. In particular, we also tried investigating the user experience with the learner GUI, more precisely, with:

1. the avatars, and its role in the learner GUI,
2. the stories, whether appealing or not for the learners,
3. the cards of characters, whether interesting or not,
4. the smart games, whether playful or too difficult,
5. the relaxing games, whether sufficiently appealing.

The focus was on identifying areas whether and which improvements should be made prior to the large scale evaluation.

In Italy, it was possible to run several sessions and, by incrementally and iteratively improving on the learner GUI prototype, it was eventually possible to gather also quantitative data where sufficient technical facilities, like a stable wireless, were available.

Task order	**Task description**
1	accessing the system via the login page
2	choosing an avatar
3	choosing a book
4	choosing a story in the spatial map of the book
5	browsing and reading the cards of characters
6	browsing and reading a story
7	browsing and playing with smart games
8	browsing and playing with relaxing games

Table 1 Usability evaluation tasks.

2.2 Experiment Participants

The evaluation in Italy counted 57 learner participants, out of which 16 are deaf, all aged 7–12, and from different locations from the North and the Centre of Italy:

- Centre of Italy: the summer school of the National Laboratories of Gran Sasso (LNGS), nearby l'Aquila; the summer school of Sacro Cuore in Avezzano; private lodging in Avezzano; a summer school for deaf children in Ciampino, nearby Rome;
  - Hearing 12 (7-9 years old); 18 (9-11 years old);
  - Deaf 1 (7-9 years old); 8 (9-11 years old).
- North of Italy: the Akademia summer school in Bolzano; the unity of audiology and phonology of the Ca' Foncello Hospital in Treviso.
  - Hearing 3 (7-9 years old); 3 (9-11 years old);
  - Deaf 5 (7-9 years old); 7 (9-11 years old).

All the participants had used at least once a PC with mouse. In the North of Italy, 5% of the children did not know the tablet and they had never used it. In each session, there were 1 or 2 children per experimenter. Among the experimenters, there were always an expert facilitator.

2.3 Experiment Tasks and Material

At the start of every session, each learner or their educators were asked some questions in order to know the learner's school class and age, and then the experimenter inserted the appropriate login information on behalf of the child. The facilitator informed all children about the goal of the evaluation, that is, to present them a system for helping their peers to better understand a story. The facilitator then asked the children to talk aloud their opinions: since the system was in its infancy, it was important that the children would tell us their opinions while using the system, on what was clear and what unclear, so as to help us improve the system with their valuable feedback.

At this point, the evaluation session started. Every child could interact with the learner GUI by using a 10" tablet or PC with mouse. During the session, the child could perform the sequence of tasks in Table 1.

In particular,

- younger children, whose age is 7–9, read either "La Vacanza Comincia" (The Holiday Begins) or "A Caccia di Delfini" (Dolphin Spotting),
- older children read either "La Mania della Competizione di Benedetto" (Ben's Racing Problem) or "Sofia e il Nano dell'Isola" (Sophia and the Island Dwarf),

and played with the associated smart games that were, on average, 1 per game level:

- 1 who game and 1 what game;
- 4 games that required reasoning about time: 1 before-after, 1 before-while, 1 after-while, 1 before-while-after;
- 2 games that required reasoning about causality: 1 cause or 1 effect, 1 cause-effect.

The learning material is reported and described in [2] and [4]; it is also available at $www.terenceproject.eu/demos$.

During the session, the appointed experimenter was observing and intervening only at critical points, when the learner definitely needed assistance. Qualitative data were thus gathered:

- via direct observations, e.g., of facial expressions, and by tracking comments per tasks;
- via indirect questions to children at critical points, e.g., if the child asks for help, if the child looks lost, if the child looks frustrated.

In the North of Italy, technical facilities, like a stable wireless, allowed for collecting reliable quantitative data through logs. The quantitative data we gathered were:

- session time, that is, the time span in between the start and the end of a session;
- for the reading task, the start and end time for reading the book selected, and the reading time per page;
- for each game instance, the number of correctly resolved game instances, and the time for their resolution before the game was over.

At the end each session, the experimenters run a debriefing phase and a short interview with indirect questions, reported in [5]. The questions were related to the usability and the experience of the learner with the system, i.e., their previous acquaintance with tablets and PCs, as well as whether they thought the story text or illustrations were appropriate for younger/older children, whether they thought the games and the interaction gestures implemented are appropriate for younger/older children, whether the avatar are nice for younger/older children, and what they would like to improve in the GUI.

2.4 Experiment Results

All the usability issues tracked, during the session or the debriefing phase, are reported in details in [5] in a specific category, per country, and should be considered for improving the design of the learner GUI. In general, a category corresponds to a task, e.g., playing with a smart game. Two categories do not correspond to specific tasks, namely, the avatar and the navigation category. Therein, we gathered issues re-occurring in different tasks, and then removed the issues from these tasks. An example is the position of the avatar during the browsing of books and stories, as well as during the reading and playing activities: positioning the avatar in the top-right corner of the screen hides the avatar, and its role in the learner GUI is too passive

or unclear. The category named “playing with smart games” is subdivided into subcategories of correlated tasks or issues, e.g., choices available to the learner, so as to facilitate the interventions of those working on the design and requirements. Table 2 briefly reports the usability results divided in positive and negative issues, where

- negative (NEG) issues if they pinpoint specific usability problems,
- and positive (POS) issues if they support design choices or purport a positive user experience.

3 Discussion and Conclusion

In light of the user-based usability evaluations, we find the following results:

- the interaction with cards needs to be improved; children did not often read the information of the cards, they often quickly looked at the images of the cards, and flickered through these; some children suggested the GUI display only the characters of each story, and not of the entire book;
- children in Italy generally liked the types of relaxing games; however, all children were frustrated when unable to play the games, due to the too fast time-out or the usability of gestures; they were also puzzled by the fact that the relaxing games were not contextualised in the learner GUI (e.g., missing points);
- playing with smart games and reading a story are tasks with the highest number of usability results, which are uniformly distributed between negative and positive; thus, they are likely to be determinant for the success of the software.

Moreover, analysing the overall results described in [5] we find that the causality games are more difficult than the time games, which may be well due to the low affordance of the software version of the causality games, as highlighted by the usability evaluations. It is remarkable that quantitative results in Italy show a correlation between age and resolution of time and causality games; in particular,

- the usability evaluations in Italy show that younger learners had more difficulties with the time and causality games than with the other games, and their highest resolution time was for the time games;
- again the usability evaluations in Italy show that deaf learners had more difficulties with the time and causality games than with the other games, and their highest resolution time was for the time games.

Such results purport that the resolution of time and causality games can give reliable indications to the adaptive engine and, more precisely, whether this can move the learners from one story level to another, as designed in [2].

Another issue, is the need of presenting instructions in a different format, clearer and more appealing. In particular, during the usability evaluations children tended not to read the instructions or these were not sufficiently clear. Moreover, some deaf children needed further assistance the first time they played with a new game level.

Category	Negative	Positive	Number negative	Number positive
Login	lack of human presence or animation	most beautiful page for many children	3.00	2.00
Navigation buttons	next button overlooked	enjoyable and usable dashboard effect	2.00	1.00
	next button area too narrow			
Avatars	no browsing besides the 2nd avatar	nice images	4.00	2.00
	no clear role of avatar after being chosen	male learners choose the male avatars, female learners choose the female avatars		
	no perception of the avatar's growth or the avatar's relations with points			
	no black avatars			
Choosing books	book titles too small	children liked the general layout	1.00	1.00
Choosing stories	too big or too poorly coloured padloacks	beautiful spatial map for choosing stories of a book	1.00	2.00
		clear spatial map for choosing stories of a book		
Reading stories	not sufficiently visible page number	deaf children in Treviso first read then looked at images for fixing in mind what they had read	5.00	3.00
	complaints of too small fonts or not nice font type	story plots were generally judged funny and creative, instructive and with a deep meaning		
	vocabulary at times too difficult	children, in particular younger ones, liked the illustration style		
	vocabulary or length of sentences too difficult for deaf children in the Centre of Italy			
	illustrations: some children complained about incoherencies between story texts and illustrations or badly resized images; some older children judged illustrations good for younger children; many deaf children complained about lack of vivid colours, and the characters being always the same or the illustrations not being realistic			
Interaction with character cards	problems with captions, e.g., too small font size or typos	older children seem more interested in reading the cards than younger children	4.00	1.00
	card under focus was not readable			
	information was not read			
	too numerous or confusing, better per story			
Playing with relaxing games	decontextualised	children generally appreciated the types of games	4.00	1.00
	frustrating due to too fast timing or gesture usability issues			
	slicing the fruit reminded of the Fruit Ninja games and, on PCs, children wanted to repeat the same gesture			
	in the toolbox games, children were expecting to be able to rotate images			
Playing with smart games			14.00	9.00
	Making a choice:		4.00	0.00
	not sufficiently usable captions, e.g., too small font size			
	more correct choices in what games			
	time and causality: allow to revise a choice			
	gray-out effects for unavailable choices not always working			
	Gesture and interaction modes:		2.00	0.00
	drag and drop slightly more difficult with mouse, but technical problems make it hard to be usable on touch-screen tablets			
	low affordance for causality games			
	Feedback:		4.00	0.00
	too fast timing			
	too prominent or not well placed yes/no feedback			
	visual metaphor of explanatory feedback for time games was not sufficiently clear for younger children			
	several children are willing to re-read a story			
	Points and instructions:		2.00	0.00
	instructions not read or not sufficiently clear			
	points not noticed or not sufficiently clear			
	User difficulties:		1.00	3.00
	some children complained about the plausibility of solutions of 3 games (1 before-while, 1 cause, 1 cause-effect)	time and causality games are more difficult than the other games		
		time and causality games are more difficult for younger learners than for older learners		
		who and what are entry level games		
	Other user satisfaction issues:		1.00	6.00
	one illustration not matching with the caption	general enjoyment of smart games, e.g., "better than the traditional boring education games"		
		liked what games' and causality games' visual metaphors or animation effects		
		younger children wowed what games		
		the majority of children from the North voted the causality games as the most beautiful		
		drag and drop were considered more challenging and appealing by the majority of learners, and older learners judged it better for them		
		children liked the visual metaphors of the feedback		
TOTALE			**38.00**	**21.00**

Table 2 Usability results, divided into negative and positive.

The usage of contextualised tutorials the first time the learners play with games may be beneficial.

According to the usability evaluations, some of the vocabulary seems at points too difficult, also at the simplest story level. However, it was found that, when prompted to infer the meaning of unknown words from the text, the deaf children generally were able to do it. All deaf children were invited to re-read the story in order to allow them to perform the games.

In conclusion all learners, hearing and deaf, were able to perform the designed tasks. Both of them preferred playing instead to reading. All of them do not really understand the role of the avatar though all learners like avatar. This is due of the fact that avatars were not very well contextualised. The difference between the two types of users is tangible in the administration of the tasks: the presence of the LIS Italian translator makes experiments more low and more time consuming. Conversely, deaf children are more critics and they are more interested to reveal us the issues to check and correct.

Acknowledgements This work was supported by the TERENCE project, funded by the EC through FP7 for RTD, ICT-2009.4.2.

References

1. A.Paramythis, S.Weibelzahl, Masthoff, J.: Layered evaluation of interactive adaptive systems: framework and formative methods. USER MODELING AND USER-ADAPTED INTERACTION (2010)
2. Arfé, B.: Repository of Stories, Deliverable D2.2. Tech. rep., TERENCE project (2012)
3. Brusilovsky, P., Karagiannidis, C., Sampson, D.: Layered evaluation of adaptive learning systems. Int. J. Cont. Engineering Education and Lifelong Learning 14(4 – 5) (2004)
4. Gennari, R.: Repository of Textual Smart Games, Deliverable D4.3. In: TERENCE project (2012)
5. Mascio, T.D., Gennari, R., Vittorini, P.: Small scale evaluation: design and results. In: TERENCE project (2012)
6. Olmsted, E., Murphy, E., Hawala, S., Ashenfelter, K.: Think-aloud protocols: A comparison of three think-aloud protocols for use in testing data-dissemination web sites for usability. In: In proc of CHI 2010: 1001 Users (2010)
7. RUSSO, J.E., JOHNSON, E.J., STEPHEN, D.L.: The validity of verbal protocols. Memory and Cognition 17(6), 759–769 (1989)
8. Slegers, K., Gennari, R.: State of the Art of Methods for the User Analysis and Description of Context of Use. Tech. Rep. D1.1, TERENCE project (2011)
9. T., D.M., Gennari, R., Vittorini, P.: Expert- based evaluation. In: TERENCE project (2012)
10. Vittorini, P.: Integrated system development, first release. In: TERENCE project (2012)