EAI Endorsed Transactions
on Future Intelligent Educational Environments

ICST.ORG

Evaluating an Intelligent Q&A System for Mobile Cultural Learning

Ioannis Doumanis[1] and Serengul Smith[2]

[1] i.doumanis@mdx.ac.uk
[2] S.Smith@mdx.ac.uk

Abstract

The paper presents a user study designed to evaluate the impact of an intelligent Question and Answering (Q&A) system for mobile cultural learning on the user's subjective impressions and retention performance. We have chosen a recent pedagogical framework for mobile learning as a theoretical foundation for this work. The framework postulates four types of mobile learning from which we have chosen to implement the second one (i.e., *High Transactional Distance and Individualised Mobile Learning (HI))*[13]. In particular, our Q&A system enables individual learners to interact with unknown cultural content under simulated mobile conditions in a well-organised and structured manner using natural language. To investigate the impact of the Q&A system on the participants' retention of cultural content, we compared two variations of the Q&A system in the lab under simulated mobile conditions. The systems differed both in terms of the approach to processing natural language (i.e., scripts vs. linguistic parsing) and style of a Q&A session (free vs. constrained). Below, we present the results of the study and a series of design recommendations that should aid in the development of more robust Q&A models for mobile cultural systems.

Keywords: assisted learning, intelligent interaction, pervasive & ubiquitous education, technology supported education, cultural information

Received on -XQH accepted on XJXVW pu blished on -XQH

doi: 10.4108/IL HHH

1. Introduction

Building effective Question & Answering (Q&A) systems that accepts free-form questions is a difficult and time-consuming process. In [1] we presented an approach for rapidly building such systems. Our approach combines an off-the-shelf Q&A authoring system called **V**irtual **P**eople **F**actory (VPF) [2] with a novel natural language matching algorithm that increases the VPF's accuracy. The Virtual People Factory (VPF) is a freely available web authoring tool designed specifically for rapid development of Q&A systems. Any developer with basic programming skills can create a conversational model online and integrate it with any desktop or web application with relative ease. However, as the VPF relies on keyword-based language processing approach and does not actually process human language, its limitations become apparent fairly quickly. For example, as the system does not recognize parts of speech (POS), it fails to distinguish keywords servicing different syntactical functions in different user input. To address this limitation, we designed and fully implemented an algorithm that adds three additional processing layers to the VPF web service that fully parse the user's input question. We also designed a dialogue manager using a Hierarchical Task Decomposition (HTD) approach [1] with the goal to enable VPF to keep the topic of a conversation even after several turns. However, the dialogue manager was not implemented and hence, it was not included in the study reported in this paper. To investigate the quality of the answers produced by the two approaches in natural language processing (script vs. parsing), we designed two and evaluated in the lab two Q&A applications (System A & System B) that provided answers to free-form questions and cultural information about a medieval castle in Greece. Each system provided users with cultural content covering popular attractions on two routes in the castle and allowed them to ask a question using plain English after each presentation was complete. To simulate the routes that users would visit during the tour each application included high quality interactive panoramic photographs (see Figure 1).

*Corresponding author. Email:jpirker@iicm.edu

Based on a short pilot experiment with three users, we formulated the following hypotheses to examine in this study:

H1: Users rate the answers returned by the parsing system better than the answers returned by the script-based system. The parsing algorithm was designed to provide more relevant answers than the scripts, thus resulting in an improved user experience.

H2: If the system does not understand the question, asking the user to rephrase a large number of times benefits retention performance, but not the overall user's experience. A moderate number (1-2) of questions is more likely to benefit both user performance and the user's experience.

H3: Providing a random answer is better than providing no answer at all. Although a random answer may not be the right answer, it may contain fragments of the information the user is seeking. This should in turn result in an increased user satisfaction as long as this process does not take too much time.

H4: Forcing participants to ask a specific number of questions per location leads to a better retention performance because it forces users to review the provided information multiple times to come up with the questions instead of when users are allowed to ask as many questions as they like per location. However, this effect (forcing participants to ask a specific number of questions per location) impacts negatively the overall user's experience with the system.

2. Related and Previous Work

A well-known system in the domain of Q&A systems is IBM's Watson [3]. Watson is a super-computer capable of analysing natural language so well and so fast that won the price of one million USD against human champions in Jeopardy. Jeopardy is an American television game show, where players have to identify clues in answers and reply appropriately in the form of questions. To achieve this level of performance Watson relies on a DeepQA architecture and ample amount of computer processing power (more than 80 teraflops – or 80 trillion operations per second [4]). The architecture inside Watson enables the system to massively evaluate in parallel multiple sources of data (e.g., natural language texts and databases) to find, synthesize, and deliver the most appropriate answers. In particular, the system generates multiple responses with an attached confidence level based on the available data. The response with the highest confidence is chosen by the system as the best answer to the user's question. To generate the responses and their confidences, the system integrates several algorithms for information retrieval, semantic analysis, automated reasoning and learning. As a result IBM views Watson not just as a Q&A system, but rather as a system for differential diagnosis [5] i.e., a system that generates a wide range of possibilities and for each develops a confidence level based on evidence from structured and unstructured data. Our approach differs in many ways. First, Watson is a proprietary system owned by IBM and by no means open to the community. IBM has recently given access to developers through the cloud to create their own "Powered by Watson applications" [6] but developers cannot access or modify the internal algorithms of the system. On a contrary, we have made our algorithm open source for the benefits of the Q&A developers' community. Then, our approach enables developers to create new knowledge in the system, simply by inserting Q&A pairs in an easy-to-use interface. Watson developers on the other hand, have to work closely with IBM to train Watson's knowledge base to match the requirements of a specific domain (e.g., medicine, shopping, etc.) [7]. Last but not least, little is known about the impact of Q&A systems with the language processing capabilities of Watson, on how users perceive the quality of their answers and how much they learn from their interactions with the system. The impact of Watson on users performing tasks collaboratively, under realistic conditions remains to be investigated.

Kuyten *et al.* [8] approach Q&A system development in a lighter way in terms of software and hardware requirements. Their system, processes natural language text input to automatically generate question-answer pairs. Those pairs are then compiled into scripts to be executed by Embodied Conversational Agents (ECAs) in health-related scenarios. The questions are presented as menus on the screen in a pre-fixed order. In an empirical evaluation of the system they compared three informed consent documents of clinical trials from the domain of colon cancer. Each of the documents was explained by an ECA using either a) text b) text or ECA performing monologue and c) text and ECA performing question-answering. The results show that none of the ECA systems resulted into better user comprehension of the documents, but users were significantly more satisfied with the Q&A version compared to the other two control conditions. We find these results to be rather expected. Although, their approach enables the full-automation of the knowledge process the constrained style of interaction most likely prevented users from asking the questions they wanted to better understand the documents. It is reasonable to assume that if users would have been allowed to ask free-form questions about the documents, their comprehension scores would have been higher.

Ada and Grace [9] are the twins Embodied Conversational Guide agents of the Museum of Science in Boston, USA. The characters use a near photo-realistic appearance and natural language interactions accompanied by a full repertoire of human gestures to engage visitors with the museum contents. Questions asked by visitors are handled by the *NPCEditor*, a component that has been made available as part of the Virtual Human Toolkit [10]. The component uses a statistical text classification algorithm that maps questions asked by visitors to the nearest question in a database with Question-Answer (Q&A) pairs. Given a large enough database, the algorithm can be effectively used in Q&A applications in limited domains [11]. The Q&A pairs can be created with ease in the tool's UI, but the process is confusing even for developers with good programming skills. We formally reviewed *NPCEditor* using the standardised cognitive walk-through technique (see [24] for

a full discussion of the process), and found a series of usability problems. A summary of our findings is as follows: First, the tool makes creating a Q&A knowledge base an unnecessarily complicated and time consuming process. It requires the user to go through a labyrinth of options where s/he has to disambiguate jargon terms almost at every step of the process. A more reasonable approach would be to enable developers to:

(i) Create an agent and adjust its properties/settings from the same panel of the editor. This will make it easier for designers to link agents with their associated properties in the system.

(ii) Create agents that have a default connection to the rest of the toolkit modules.

(iii) Define the states of the dialogue as an agent property (e.g., "states").

(iv) Set the initial state of the dialogue from within the above property.

(v) Access important agent properties and their background code by default. For example, the "Type" property used to handle off-topic responses, and a property needed to receive messages from the computer vision module, should be made available with the toolkit installation.

Second, the process by which the system learns the question-answer mapping is perhaps the biggest problem of the toolkit. It uses a statistical text classifier for mapping questions to question-answer pairs in the database. Designers can tune several of the classifier parameters, assuming of course they have the necessary knowledge to do so. The training process should take place in real-time, during the entering of the question-answer pairs in the database. Unless explicitly requested by the user, such advanced functions should be hidden and fully automated.

An evaluation of Ada & Grace [9] with actual visitors of the museum showed that this natural language processing approach was robust enough to handle well utterances that appear in the classifier training data (known utterances) and those, that were not in the training data (unknown utterances). The robust NLP coupled with the high-quality avatars of the twins created a "jaw-dropping" experience for visitors. However, the study did not evaluate the impact of the statistical natural language approach on what the visitors actually learnt from their interactions with the twins. We have evaluated the impact of a similar statistical language processing approach as opposed to a linguistic approach on the participants' retention of cultural content.

The *NPCEditor* was also used an implementation of an autonomous tour guide for the American Military in Second Life [17]. The agent (named staff duty office Maleno) watches two virtual islands in Second Life and provides information about the US army. Visitors to the islands can also participate in activities such as a quiz, a helicopter ride and a parachute jump. Apart from the usability problems (see Appendix A) Q&A authors may face with the NPCEditor, the full evaluation of the user's experience with the tour guide agent is yet to be conducted.

The Dubrovnik city tour guide agent [18] is an Embodied Conversational Agent (ECA) capable of adapting its behaviours to match the user's cultural background. In the current implementation, the ECA can dynamically change its verbal and non-verbal behaviour to match the American, Japanese and Croatian styles of communication. The verbal interaction is specified in AIML (Artificial Intelligence Modelling Language) scripts. As discussed in [1], AIML is not designed to process human language and to understand the meaning and structure of words and sentences. Hence, any additional piece of knowledge (including variations of the same question) will have to be written in AIML scripts. This makes creating a Q&A database a time-consuming and complex process. Finally, the impact of such an ECA to the experience of a visitor to the city of Dubrovnik is yet to be evaluated.

Another project that has explored the use of simple Q&A in a tour guide of a Virtual Environment is the REal and Virtual Engagement in Realistic Immersive Environments (REVERIE) [19] [20]. The platform features both human-to-human and human-to-agent interaction. Users can adapt the basic features of their avatars and can also create photorealistic 3D representations of them to interact with other human-users and ECAs in virtual environments (VEs). The interaction is controlled by the Microsoft Kinect sensor* that is used to animate an avatar in real time or to create a photorealistic replica of the user in the virtual world. Once of the virtual environments the project has implemented is a multi-user 3D virtual representation of the plenary chamber of the EU parliament in Brussels. Users of the environment can take part in a guided tour of the parliament (given by an autonomous ECA) and participate in open debates on various topics (e.g., on Multicultural London) with other users from around the world. The tour guide agent walks users around the parliament while providing information in both verbal and non-verbal form. The agent's reasoning framework can process and respond to basic verbal and non-verbal input from the user. For example, when users enter the virtual environment they can answer accordingly to the agent's request to start the tour by using either a head nod or say yes or a head shake or say no. An expert evaluation of the environment found that by merely listening to the amount of information the agent presents during the tour, will not hold the user's attention. Users should be able to ask some basic comprehension questions either during or at the end of the tour.

3. Theory of Mobile Learning

A theory of mobile learning is essential when investigating the impact of intelligent Q&A systems on users of mobile cultural systems. Moore's transactional distance [TD] theory postulates that there is a cognitive distance between instructors and learners in an educational setting. This distance is *"a psychological and*

* http://www.microsoft.com/en-us/kinectforwindows/

communication space to be crossed, a space of potential misunderstanding between the inputs of instructor and those of the learner" [12]. The theory provides three components that have to work in synergy in order to control and manage the transactional distance: (1) communication or dialogue between teachers and learners; (2) the structure of the provided curriculum and (3) the role of the learner autonomy or self-directness in deciding what, how and how much to learn. Park [13] adapted the theory to incorporate mobile learning. The pedagogical framework of mobile learning he developed categorizes mobile learning into four types (see Figure 1):

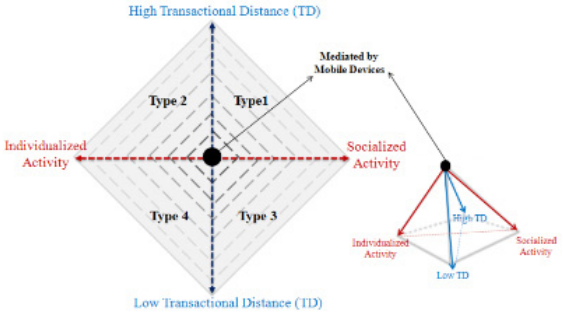

Figure 1. Four types of mobile learning a pedagogical framework

(i) **High Transactional Distance and Socialised Mobile Learning (HS).** A mobile learning activity can be classified as this type when 1) there is a high transactional distance between a group of learners and their instructor or institutional support; 2) there is high degree of learners' communication, negotiation and collaboration; 3) the group learning materials or rules of activity are predetermined and are delivered through mobile devices; 4) transactions occur mainly among learners involved in group learning or projects with minimal instructor intervention (e.g., to facilitate the group activity)

(ii) **High Transactional Distance and Individualised Mobile Learning (HI).** A mobile learning activity can be classified as this type when: 1) there is a high transactional distance between the individual learner and his/her instructor or institutional support; 2) the learning materials and resources are tightly structured and well organised (e.g., recorded lectures, tutorials) and are delivered to individual learners through mobile devices; 3) the individual learners are self-directed enough to control their learning process in order to master it; 4) the transactions mainly occur between the learner and the content.

(iii) **Low Transactional Distance and Socialised Mobile Learning (LI).** In this type of mobile learning individual learners interact with each other and the instructor using mobile devices. They have: 1) less transactional distance with the instructor or institutional support; 2) loosely structured learning materials and resources; but 3) work together with other learners in order to achieve a common goal (e.g., to deliver a project) and 4) naturally engage in social interactions (e.g., negotiations, frequent communication, etc.)

(iv) **Low Transactional Distance and Individualised Mobile Learning (LI).** This last type of mobile learning refers to 1) less transactional distance between the instructor and the learner; 2) loosely structured and undefined learning materials and resources; 3) the learners interacting directly with the instructor and 4) the instructor leads and controls the learning process in an effort to meet individual needs of the learners while maintaining their independence.

We have chosen to implement an intelligent Q&A system for cultural learning based on the second type of mobile learning (i.e., **High Transactional Distance and Individualised Mobile Learning (HI))** for two reasons:

a) We wanted to investigate the learning needs of specific type of visitors in archaeological attractions, those who visit an attraction for the first time and for whom retention of cultural content is important (e.g., students with an interest in culture, cultural experts, a history hobbyist, educators, etc.).

b) We did not have access to a big pool of participants that would allow us to implement and evaluate the socialised models of mobile learning.

Based on the HI model of mobile learning we designed a system that: 1) Required participants to retain cultural information about locations they had never visited before; 2) Organised the cultural content into smaller parts, each narrated by a high-quality Text-to-Speech Engine (T2S) in a simulated mobile environment. In addition, users of the system could interact with the content using free-form questions; 3) we used final year undergraduate students at the University of Middlesex with interest in cultural content. Undergraduate students at Middlesex University are taught all the necessary skills to master their own learning from year one [14] [16]. Therefore, we have assumed that the self-directness of the participants was sufficient to control their learning of the cultural content with the Q&A system; 4) Students experienced and interrogated the cultural content only through the use of the Q&A system, and did not interact with other students or an instructor.

4. Prototype Systems

For this experiment, we designed two simple interfaces: "System A" and "System B" (see Figure 2) using Microsoft Visual Studio 2010. Each system provided participants with cultural content covering popular attractions on two routes

in the castle and allowed them to ask questions using plain English after each presentation was complete. Each route included three locations to visit in turn (labelled Locations A-C and Location D-F). The systems utilized either the script-based or a parsing-based approach to process natural language questions. The script-based system is based on a third-party system, called the **V**irtual **P**eople **F**actory [2] while the parsing approach on our own algorithm for processing natural language questions based on Antelope's (Advanced Object Oriented Processing Environment) NLP framework [21]. An expert human-guide wrote the presentations and crafted the initial conversation corpus using the **V**irtual **P**eople **F**actory (VPF) authoring tool [2]. The interface of both systems is simple enough to use without any previous training and it is divided into the following sections:

(i) **The System section:** This section features an input field for typing a question, an output field for displaying the system's response, a drop-down menu for defining the location the user is visiting, and two buttons for controlling the speech output of the system. The synthetic speech was generated by a T2S engine.

(ii) **The Indicators section:** This section provides information about the total number of questions asked, the database question the system matched the input question to, and the part of the presentation where the user is currently listening.

A simple key combination activates the "Castle Window" that displays an interactive 360° panoramic representation of each location participants had to visit in the castle (right side of Figure 2). In case of an unknown input, i.e., the participant asked a question that the system failed to match with the database, the system requested the participant to rephrase the question. If the participant failed to rephrase the question in a way the system could understand a specific number of times (different for each location), the system returned a random answer from the database. This was done to investigate the impact of varied number of times participants had to rephrase a question on their retention performance and experience with the prototypes.

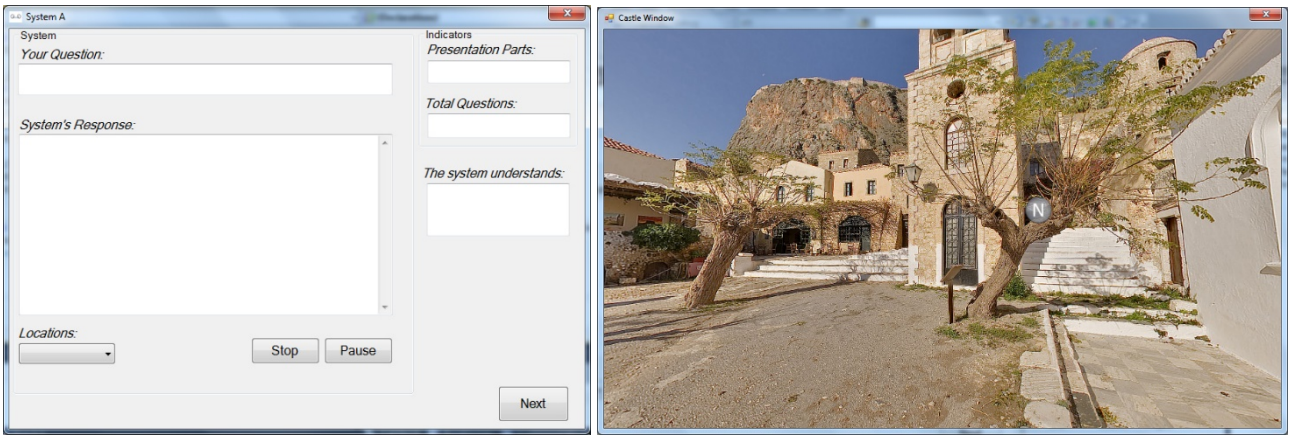

Figure 2. One of the two prototype systems with the panoramic window

5. User Study

To test our initial hypotheses (see H1 to H4) we designed a study using the A/B method [23], where the participants had to compare alternative versions of two tour guide systems. This method was deemed as necessary as we wanted to evaluate the impact of each prototype (featuring different approaches for natural language question & answering (Q&A) and style of Q&A), on the users' subjective experience and retention performance.

Population: Our approach was entirely user-driven. We initially asked a group of three users to test a preliminary prototype and tell us their requirements. Based on their responses, we formulated a number of hypotheses to test refined the original prototype and evaluated it with a group of twelve (12) new participants. All participants were undergraduate students of Middlesex University who participated for course credit and were randomly assigned to conditions. None of the participants was a local resident of the castle or had visited the castle before. The participants had a variety of computer science majors and computer experience backgrounds and they were all native speakers of English.

Task: To ensure that the systems would run properly, participants interacted with the systems using a high-end

laptop (i.e., the Sony Vaio FZ21Z model[†]). After the experimenter provided a brief explanation about the purpose of the experiment, the participants began the task, which was to uncover information about six locations of the medieval castle and ask questions after the completion of each presentation. They were asked to perform this task once using System A and again using System B. To make it easier for participants to understand the provided information, each presentation was divided into parts and an interactive panoramic representation of each location was integrated in the systems. Participants could interact with the panoramic while listening to a presentation, thus relating the provided information to the actual locations. Half of the participants in each group were told to ask as many questions as they liked per location as long as the total number was not greater than twelve. The other half was restricted to four questions per location. In case the system failed to process one of the questions, participants were asked to rephrase their question as many times as necessary, until they got an answer. Once the system provided an answer, participants were asked to rate thirteen statements on a 10-point scale (1 = no answer 10 = perfect answer). Examples of these statements are clarity and wording of the answers. After the participants had visited all locations, they were asked to write down what they could remember from the presentations (and answers) about each location in total, and freely comment on their overall experience with the systems.

Conditions: The independent variables in this experiment were:

- The approach used for question processing (scripts vs. parsing),
- Style of Q&A (forced vs. free), and
- Order of task (first route then second vs. vice versa)

As dependent variables we measured:

- Performance (i.e., percentage of propositions recalled from the content), and
- The ratings in the subjective impressions questionnaires.

The variable language processing method was manipulated within-subjects (see Table 1), whereas the order of task between-subjects. Participants were randomly assigned to the four experimental conditions: 1) script-based system with the first route (i.e., Locations A-C) vs. parsing-based systems with the second route (i.e., Locations D – F) or 2) script-based system with the second route (i.e., Locations D-F) vs. parsing based system with the first route (Locations A-C).

[†] http://www.laptopsdirect.co.uk/Sony__VAIO_FZ21Z_VGN-FZ21Z/version.asp

Measures and methods: The only objective variable that was used in this experiment was the accuracy of the answers to the free recall test. The subjective measures were the responses to the items of the questionnaire. The questionnaire items used a 10 point scale (1= no answer 10 = a perfect answer) to measure the subjective impression of the participants of the answers provided by the systems.

Table 1. The experimental design

Participants N=12	Script-based system	Parsing-based system
1-6	*First Route* Subjective impressions/Free recall test/No. Of Questions	*Second Route* Subjective impressions/Free recall test/No. Of Questions
7-12	*Second Route* Subjective impressions/Free recall test/No. Of Questions	*First Route* Subjective impressions/Free recall test/No. Of Questions

Our choice of 10-point scale was consistent with that done in other similar studies [15]. The questionnaire addressed several dimensions of the subjective impressions of the answers such as clarity, sense, fun, etc. (see Table 3 for the full list of items)

6. Results and Discussion

Although we recognise that the sample size used in the study is small (12 participants), and any significant results are less than "statistically valid", it does not mean that they are "less than valid". In fact, according to the applied user research literature [22] [23], small sample sizes generate viable results when it comes to large differences between designs or to discovering common usability problems. Therefore the results reported below should be considered with caution for their generalizability and depth as more research is needed with a bigger size subject pool.

Key Performance Findings: We measured the total number of concepts recalled from the presentations. As a concept we defined one or more sentences that cover the same topic. For example, the following sentence *"The facing on the main gate, like the moulding on the wall, and the corbeling of a small bartizan located to the upper right of the portal, are all made of porous rock, quarried nearby",* is one concept that covers the material by which the main gate of the castle was constructed. Each test was scored as a percentage of the correctly reproduced

concepts. Table 2, shows the overall participants' retention performance

Table 2. Mean retention performances

Order Of Task	System	Mean	SD
First route vs. Second Route	Scripts	15.8	16.4
	Parsing	10.8	10.7
Second Route vs. First Route	Scripts	6.6	4.8
	Parsing	4.5	3.8

This interaction was further analysed using simple main effects analysis. It showed that the variation of order of task significantly influenced the participants who were forced to ask four questions per location ($F(1, 20) = 10.805$; $p < .01$) but not the participants who were allowed to freely ask questions (see Figure 3).

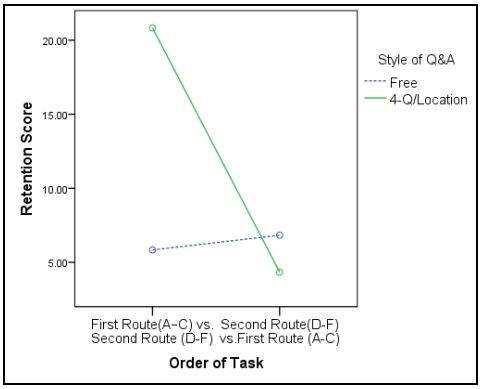

Figure 3. The interaction of retention score for order of task and Q&A style

A close inspection of the descriptive statistics (see Table 3), revealed that the participants who were forced to ask four questions per location, performed better overall in the first order (mean A-C/D-F = 20.8), than in the second order (mean D-F/A-C = 4.3). This effect is independent of the type of system used (parsing or scripts). The participants who used the script-based system performed better in the first route (mean A – C = 24.6) than in the second route (mean D – F = 5.3). The participants who used the parsing-based system performed vice versa (mean A – C = 3.3 vs. mean D – F = 17).

Table 3. Retention performance as a function of Q&A style and order of task

Q&A System	Order of task	Questions	Mean	SD
System A Scripts	First Route (Location A – C)	Free	7.0	4.5
		4-Q/ Location	24.6	20.4
	Second Route (Location D – F)	Free	8.0	5.2
		4-Q/ Location	5.3	5.0
System B Parsing	Second Route (Location D – F)	Free	4.6	5.0
		4-Q/ Location	17.0	12.2
	First Route (Location A – C)	Free	5.6	4.5
		4-Q/ Location	3.3	3.5
Total	First Route vs. Second Route	Free	5.8	4.4
		4-Q/ Location	20.8	15.6
	Second Route vs. First Route	Free	6.8	4.5
		4-Q/ Location	4.3	4.0

This finding suggests a correlation between the content participants experienced in each route and the type of question-processing that was used. The content that participants experienced in the second route was domain-specific (i.e., about churches), while in the first segment it was open-ended (i.e., a variation of attractions). Therefore, parsing is a better approach for processing more domain-oriented questions than scripts, while scripts are a better approach for processing more open-ended questions than parsing.

Additional Performance Findings: Although there is no significant effect of the system type on the retention scores, it is clear from Table 2 that participants performed on average better with the script-based system. As scripts were more accurate [1], participants got better answers to their questions than when using the parsing system. For every unknown input the system would ask the participant to rephrase the question. This means that the parsing-based system would ask the participants to rephrase an unknown question, more times than the script-based system. We observed in the lab that this annoyed them and most likely distracted them from the information they already had in their minds about the locations. Therefore, the first part of our hypothesis (see H2) is invalid as asking participants to rephrase a question a large number of times does not lead to an enhanced retention performance or improved user experience. Then, based on the participants' comments we argue that the second part

of our hypothesis (see H2) is most likely valid, i.e., asking a user to rephrase a question once could lead to improvements in both retention performance and the user's experience. However, apart from the participants' comments, we do not have any other evidence to fully support this claim. In relation to the style of question-asking, the table below (see Table 4) shows that the participants who were forced to ask four questions per location performed better overall (mean 4Q/Location = 12.5) than those who were allowed to ask as many questions they liked per location (mean free = 6.3). However, a one-way ANOVA, testing the difference between the means of both styles failed to reach significant levels ($p > .05$). The lack of significance can be attributed to the small number of participants in each group (6 participants / group). Additional research is needed to statistically validate the users' tendency to remember more information (as evident in the descriptive statistics) when they are forced to ask a specific number of questions per location, than when they are allowed to freely ask questions.

Table 4. Constrained/Free question asking per system

Style of Q&A	System	Mean	SD
Free	Scripts	7.5	4.46
	Parsing	5.1	4.3
4Q/Location	Script	15	17
	Parsing	10.1	11.01

Furthermore, in the lab we observed that those participants got frustrated from having to review the content several times in order to come up with the specific number of questions. Both findings provide grounds that my hypothesis (see H4) could be valid and that forcing participants to ask a specific number of questions enhances retention performance, but not the overall user's experience.

Subjective Assessment: Table 5, shows the mean responses for the questionnaire items for the different system and order of task conditions. The questionnaire was highly reliable (Chronbach's a= 0.89). The participants rated all the items of the questionnaire almost similarly. Therefore, my hypothesis (see H1) that the parsing system improves the user's experience by providing more relevant answers is not rejected. Except for "fun" and "accuracy", participants seem to have perceived both methods for processing natural language questions similarly. We performed an ANOVA taking the participants' ratings for each of the questionnaire items as a dependent variable, and type of system and order of task as independent variables. It showed a statistically significant effect of order of task on the following questionnaire items:

- Item 6 *("Fun")* ($F(1, 20) = 4.616$; $p < .05$)
- Item 8 *("Interesting")* ($F(1, 20) = 6.943$; $p < .05$)
- Item 11 *("Tiresome")* ($F(1, 20) = 12.454$; $p < .01$)

All effects are clearly because of the variation of content across the order conditions. Participants in the first order, experienced content from the first route (i.e., Locations A – C) with the script system, then content from the second route (i.e., Locations D – F) with the parsing system, while participants in the second order experienced the content vice versa.

Table 5. Mean responses to the questionnaire items

Measures	Order 1		Order 2	
	Scripts/ Parsing	SD	Scripts/ Parsing	SD
Fun	5.8/ 5.0	1.2/ 1.7	6.5/ 6.7	1.3/ 1.3
Interesting	5.9/ 4.7	1.0/ 1.7	6.5/ 6.6	1.2/ 0.5
Tiresome	2.3/ 2.5	0.9/ 1.1	4.0/ 4.1	1.3/ 1.3
Clarity	6.8/ 6.1	2.7/ 1.9	6.6/ 6.7	1.7/ 1.0
Wording	6.5/ 6.1	2.5/ 1.7	6.2/ 6.8	1.7/ 1.0
Sense	6.3/ 5.8	2.2/ 1.9	6.0/ 6.8	1.8/ 0.4
Understandable	6.8 / 6.3	2.2/ 1.6	6.5/ 7.1	2.0/ 0.9
Simplicity	6.6 / 6.8	1.0/ 1.4	6.6/ 6.7	1.5/ 0.8
Annoying	2.2/ 2.9	0.9/ 1.7	2.4/ 2.3	1.3/ 0.9
Intelligent	6.1/ 5.1	1.9/ 1.7	7.0/ 6.6	1.4/ 0.6
Stimulating	5.2/ 4.5	1.6/ 1.8	5.8/ 5.9	0.6/ 0.6
Tiresome	2.3/ 2.5	0.9/ 1.1	4.0/ 4.1	1.3/ 1.3
Unpleasant	2.0/ 2.2	0.7/ 1.5	2.9/ 2.8	1.2/ 1.5
Accuracy	6.1/ 5.2	2.4/ 1.9	6.5/ 7.0	1.9/ 1.1

Comments: After participants wrote down what they could remember from the presentations, they were asked to write down openly what they think about the two

systems. From the comments participants made, we selected the following and grouped them accordingly.

System A Design (Scripts):

1) Simple and easy to use, with surprisingly accurate answers.
2) Faster than system B

System B Design (Parser):

1) Too slow (it takes up to a minute to load)
2) One participant said that he did not find the answers he was looking for, while another said that this system is more accurate.

The above comments about the two systems are consistent with the patterns of questionnaire scores (see Table 5). The script-based system was generally perceived by the participants as faster and more accurate, than the parsing-based system.

General improvements (both systems):

1) If the system cannot answer at least one of the questions, it should take the user back to the same paragraph s/he was reading.
2) Both systems should use easy vocabulary and clearer sentence-structure.
3) When the user enters a question, provide suggestions, like Google, to help the user to ask the correct question.
4) If a question cannot be answered, at least the second time, the system should take the user back to the same paragraph s/he was reading.
5) The speed of the text-to-speech (T2S) should be slower.

Participants provided a plethora of suggestions for improvements that can radically enhance the overall user's experience. An improvement of particular importance is the number of times the system should ask the participant to repeat the question, and the systems' action afterwards. Participants suggested that this should happen just once. The second time, the system should take the participants back to the content it was narrating. This comment provides an indication that returning a random answer (see H3 hypothesis) when the system fails to interpret a question may not be a good idea at all. However, as there is insufficient evidence to support this claim, this issue needs to be investigated further in future experiments.

7. Design Guidelines

On the basis of the findings of our empirical study, we established a number of design guidelines. The guidelines are based on quantitative and qualitative evidence generated from the experiment and should aid developers in the design of more robust Q&A systems for mobile learning. In total, we identified four recommendations that are presented below in layman terms:

Consider the four types of mobile learning and design an intelligent Q&A system according to your needs. We have found that the four types of mobile learning in the Park's pedagogical framework [13] provide sufficient guidance for the design of intelligent Q&A systems for cultural learning that match varied learning needs. In the empirical work reported in this paper, we investigated the high transactional type of mobile learning (for reasons explained above) and implemented a system based on the individualised approach. However, given the overall low retention performances of the participants in our study (see Table 2), future work could investigate the social approach in mobile learning using a modified version of the Q&A system.

Prevent the system from repeatedly requesting users to rephrase an unknown to the system question. Based on the participants' reactions we observed in the lab, it is clear that they were distracted by the repeated requests of the systems to rephrase an unknown question. As this has impacted their retention of the content (see Performance Measures for a discussion), consider either to: a) provide a random answer to the query or b) take the user back to the narration of the content and ask him/her to review it again.

Have a human guide to create the cultural content and craft the Q&A pairs for the system. The results of this study suggest that users perceived the cultural content (i.e., presentations about the locations of the castle and answers to their questions) of the systems positively (see Table 5). Therefore, to ensure positive user experiences of the content, it should be created by an expert human guide of the archaeological attraction and not based on a guide book.

Do not limit the number of questions per location of a tour. We found a strong indication that when users are forced to ask a specific number of questions per main narration of a tour, their retention performance increases. However, we also observed that their perception of the friendliness of the system decreases. Therefore, unless the desired outcome of the interaction with the mobile guide system is enhanced retention performance, allow participants to ask as many questions as they like per location of a tour.

8. Conclusion and Future Work

This empirical study provided evidence that retention of cultural heritage content is related to the accuracy of the method used in the question-answering session with the system. The more robust the method is, the less are the chances for participants to become distracted and forget what they learnt from the interaction with the system. In our previous work [1] we compared the two approaches for natural language processing (scripts vs. parsing) used in the prototypes participants used in the study. We found

that scripts are overall the more robust approach. In the current study, however, we found evidence that parsing is better for processing more domain-oriented questions (e.g., the architecture of churches) than scripts. We plan to optimise the speed and accuracy of parsing by: (a) use a more recent version of Antelope's (Advanced Object Oriented Processing Environment) NLP framework [21] and (b) optimise our search and match algorithm by using an SQL database (instead of a flat XML file) to store and retrieve data. With these improvements in place, parsing may prove to be superior across all evaluation metrics, if the study reported in this paper is repeated.

Then, we found a strong indication that when retention performance is the desired output of the interaction process with a Q&A system, participants should be required to ask a specific number of questions per location. However, this approach is frustrating for users as it forces them to review the content many times in order to come up with the required number of questions. A last important finding has to do with the requests to rephrase a question when the system fails to match it. We found that the repetitive requests annoy participants and affect their retention performance. Therefore, to ensure an optimal Q&A session, the request should be repeated just once, as participants suggested, or the repeat messages should be built in a way to allow users to figure out how to ask the system questions to avoid improper responses. Future work will investigate the socialised approach in mobile learning on the same type of unknown cultural content (see *High Transactional Distance and Socialised Mobile Learning (HS)*). We find this type of mobile learning particularly appealing. This is because participants can explore the cultural content with the help of the Q&A system, but also by getting help from other users that experience the same content in the same location. When the Q&A fails to answer a question appropriately, it could motivate participants to connect to the closest user to their location and ask the question (e.g., through text or voice). We believe that such socialised approach could maximize the utility of the Q&A system as a tool to experience unknown cultural content in archaeological attractions.

References

[1] Doumanis, I., Serengul, S., "Beyond AIML: Rapid Prototyping of a Q&A system" IN: proceedings of the 9th International Conference Intelligent Environments (IE'13) 16-17 July 2013, Athens Greece, pp. 530-540

[2] Virtual People Factory (VPF) (2013) (Available from: http://www. virtualpeoplefactory.com) [Accessed June 02 2014].

[3] What is Watson? (Available from: http://www.ibm.com/smarterplanet/us/en/ibmwatson/) [Accessed June 02 2014].

[4] Deedrick, T., "It's Technical Dear Watson" accessed 02 June 2014, (http://www.ibmsystemsmag.com/ibmi/trends/whatsnew/It%E2%80%99s-Technical,-Dear-Watson/)

[5] The DeepQA Research Team (Available from: http://researcher.watson.ibm.com/researcher/view_group_subpage.php?id=2159) [Accessed June 02 2014]

[6] IBM Watson Ecosystems (Available from: https://developer.ibm.com/watson/docs/ibm-watson-ecosystem/overview/) [Accessed June 02 2014]

[7] Ferruci, F., Brown, E., Chu-Carroll, J., David, G., Kalyanprur, A, A., Lally, A., Murdock, A., Nyberg, E., Prager, J., Schlaefer, N., and Welty C., (2010): "Building Watson: An Overview of the DeepQA Project" AI Magazine Vol 31, No 3, accessed 02 June 2014 from: http://www.aaai.org/Magazine/Watson/watson.php

[8] Pascal., K, Bickmore, T., Stoyanchev, S., Piwek, P., Prendinger, I., Mitsuru, I., (2012): "Fully Automated Generation of Question-Answer Pairs for Scripted Virtual Instruction" Proceedings of the 12th International Conference on Intelligent Virtual Agents (IVA), Santa Cruz, CA, USA, September, 12-14, 2012, Springer: pp: 1-14

[9] Swartout, W., Traum, D., Artstein, R., Noren, D., et. Al (2010): "Ada and Grace: Toward Realistic and Engaging Virtual Museum Guides". 10th Int. Conf. on Intell. Virtual Agents, vol. 6353, Springer, Philadelphia, PA (2010), pp. 286-300

[10] Institute for Creative Technologies (ICT), "The ICT Virtual Human Toolkit", version 0.9.42.514, computer program and manual, University of Southern California, California, USA.

[11] Leuski, A., Traum, R. D., "NPCEditor: Creating Virtual Human Dialogue Using information Retrieval Techniques" AI Magazine Vol 32, No 2 accessed 02 June 2014 from: http://www.aaai.org/ojs/index.php/aimagazine/article/view/2347

[12] Moore, M. G., (1997): "Theory of transactional distance". Keegan D., ed. In "Theoretical Principles of Distance Education" Routledge, pp. 22-38

[13] Park, Y. (2011): "A Pedagogical Framework for Mobile Learning: Categorizing Educational Application of Mobile Technologies Into Four Types. The International Review of Research in Open and Distance Learning, Vol 12, No 2 accessed 02 June 2014 from: http://www.irrodl.org/index.php/ irrodl/article/view/791/1699*

[14] Smith, S. and Edwards, J. A. (2012): "Embedding information literacy skills as employability attributes". ALISS Quarterly, 7 (4). pp. 22-27, ISSN 1747-9258.

[15] Stevens A, Hernandez J, Johnsen K, Dickerson R, Raij A, Jackson J, Min Shin, Cendan JC, Duerson M, Lok B, Lind DS (2006): "The use of virtual patients to teach medical students communication skills". The American Journal of Surgery, Volume 191, Issue 6, pp. 806-811, June 2006

[16] Bernaschina P. and Smith S. (2012): "Embedded writing instruction in the first year curriculum", Journal of Learning Development in Higher Education: Special Edition: Developing Writing in STEM Disciplines, The Association for Learning Development in Higher Education, November 2012, ISSN: 1759-667X

[17] Jan, D., Roque, A., Leuski, A., Morie, J., Traum, D. (2009): "A virtual tour guide for virtual worlds." In: Ruttkay, Z., Kipp, M., Nijholt, A., Vilhj´almsson, H.H. (eds.) IVA 2009. LNCS, vol. 5773, pp. 372–378. Springer, Heidelberg (2009)

[18] Huang H.H, Kateryna T., Toyoaki N., Cerekovic A., Vjekoslav L., Goranka Z., Igor S. P, and Yukiko Nakano (2008): "An Agent Based Multicultural Tour Guide System with Nonverbal User Interface, the International Journal on Multimodal Interfaces" Vol. 1 No. 1, pp 41-48, Springer Press, April 2008.

[19] L. Argyriou, P. Klavdianos, J. Wall, M.F. Rivera, N. Hajimirza, K. Apostolakis, P. Daras, G. Kordelas, T. Tsiodras, T. Zahariadis, F. Kuijk, C. Stevens, V.S. Broeck, E. Quacchio, S. Poulakos, and A. Marra, "D3.2 Report on design specification of graphical user interface for use case 1". http://www.reveriefp7.eu/resources/deliverables/.

[20] S. Poulakos, P. Klavdianos, L. Argyriou, K. Apostolakis, D. Alexiadis, J. Wall, J. de Vet, P. Fechteler, and M. Bakker, M., "D3.2 Report on design specification of graphical user interface for use case 2". http://www.reveriefp7.eu/resources/deliverables/.

[21] Proxem Advanced Natural Language Object oriented Processing Environment., version 0.8.7 computer program and manual, Proxem, France.

[22] Sauro, J. (2014). Nine Misconceptions about Statistics And Usability. [Blog] Measuring Usability. Available at: http://www.measuringusability.com/blog/stats-usability-errors.php [Accessed 1 Sep. 2014].

[23] Sauro, J. and Lewis, J. (2012). Quantifying the user experience. 1st ed. Amsterdam: Elsevier/Morgan Kaufmann.

[24] Doumanis, I. (2013): "Evaluating Humanoid Embodied Conversational Agents in Mobile Guide Applications". Thesis (PhD). Middlesex University