

A Usability Study of the GIVE-2 Environment

Nikiforos Karamanis

Ielka van der Sluis
Trinity College Dublin
Ireland

Gavin Doherty

name.surname@cs.tcd.ie

Abstract

A usability study applied to the GIVE-2 environment unveiled issues which were passed on to the organisers of the challenge and may have otherwise impacted on the outcomes of GIVE-2. The study also suggests that interacting with GIVE-2 is perceived as performing a task rather than playing a game and that prior gaming experience has an effect on task completion.

1 Introduction

During the first Challenge on Generating Instructions in Virtual Environments (GIVE-1), five Natural Language Generation (NLG) engines were evaluated over the Internet using data from 1143 trials. This made GIVE-1 the largest NLG evaluation effort to that date (Byron et al., 2009) and motivated a subsequent run in 2010 (GIVE-2).

A GIVE participant receives instructions generated by an NLG engine which are meant to help her find a trophy in a virtual environment. So, in addition to facilitating the large-scale evaluation of various NLG engines, the GIVE environment can also be regarded as an interactive application which may be subject to usability evaluation.

Performing usability tests iteratively during the development of an interactive application has become common practice in the field of Human-Computer Interaction (HCI). Usability tests are often formative, aiming to unveil issues that may not always be obvious to developers by getting feedback from users. Qualitative techniques such as direct observation and interviews with users are frequently employed to elicit this feedback.

We are not aware of research reporting on formative qualitative usability testing during the development of NLG-driven applications. This paper presents such a study applied to the GIVE-2 environment. The study uncovered usability issues that were passed on to the organisers of the

challenge and may have otherwise impacted on the results of GIVE-2. The study also suggests that interacting with GIVE-2 is perceived as performing a task rather than playing a game and that prior gaming experience has an effect on task completion and the strategies that the participants adopt. These wider issues are discussed in relation to the purpose and the generalisability of the challenge.

2 The GIVE environment

In the GIVE environment, a player can move around and manipulate objects. To finish the game successfully, she has to find a trophy which can be accessed by pressing a sequence of virtual buttons. The environment includes buttons which do nothing or activate an alarm tile. Stepping on this tile causes the player to lose the game unless she deactivates it first by pressing another button.

The NLG engine has access to the current state of the world and to an automatically computed plan of the actions that the player should perform to find the trophy. When the player performs an action, the engine is notified and may generate a new instruction expressing the next step in the plan or the first step in a new plan.

The virtual environment in GIVE-1 was made of tiles and the participant could only move from one tile to the other following stepwise instructions of the form “move three steps forward”. By contrast, GIVE-2 permits continuous movements, which makes the generation task more challenging. The inclusion of distractor buttons requires appropriate reference to virtual objects which is facilitated by landmarks such as chairs.

The five NLG engines which participated in GIVE-1 were compared using both objective (task success, completion time, etc) and subjective measures (task difficulty, instruction clarity, helpfulness, etc). The approach was mostly quantitative relying on log data for the objective measures and by asking participants to fill in an online question-

naire after each trial. For many of these measures only the data from successfully completed games was taken into account (61% of all trials). The results of the online evaluation were validated by comparing them with the results of a laboratory experiment (Koller et al., 2009).

3 Formative qualitative usability testing

GIVE-1 was mostly focused on comparing NLG engines. However, since the GIVE environment serves as the main vehicle of the evaluation, investigating its usability appears to be relevant too. One way to do so is by following Nielsen's "discount usability" approach (Nielsen, 1989). This approach relies on a small number of participants and has become popular for evaluating the usability of web applications.

The underlying idea is that just a few (typically 3-5) participants suffice to discover the most outstanding usability issues at a certain stage of development. After these issues have been addressed, another small test points to additional issues. So performing formative tests with a few participants in many iterations is thought to unveil a larger number and variety of issues than a single test with more participants (Krug, 2006, pp.135-139).

A formative usability test is best seen as an attempt to unveil issues that may be less obvious to the developers of an application. Thus, it is distinct from a laboratory study such as the one conducted by Koller et al. (2009). Additionally, a usability study pays particular attention to participants who have been unable to complete the task.

Due to the small number of participants, a formative test often relies on qualitative techniques. Interviews give participants the opportunity to elaborate on answers that they gave in a questionnaire and even extend the discussion to aspects of the interaction that may not be included in the questionnaire. Direct observation may capture other unanticipated issues with the application.

4 Study set-up

Participants: 5 male participants took part in the study as unpaid volunteers. All were graduate students in our University. Three were between 20-29 and two between 30-39 years old. Two were native English speakers born in Ireland. Two others described their command of English as "near-native" and the fifth as "very good".

The participants' gaming experience varied.

Three used to play games frequently until recently. One (S1) was still playing for about 20 hours per week. One (S4) did not play games at all.

Method: The study took place several months before the official run of GIVE-2 in a usability laboratory which provided for screen capturing, keyboard logging and video recording of the sessions. The participants interacted with the GIVE-2 environment on a desktop computer. The environment was accessed via the GIVE website¹ and utilised a proof-of-concept NLG engine. This seemed appropriate since the study investigated the usability of the environment and not the quality of the generated instructions.

A questionnaire consisting of 35 items, grouped to cover 8 themes (helpfulness, understandability, affective dimension and naturalness of the instructions, timing and amount of information, immersiveness and technical problems), was presented to the participants on paper. The questionnaire was an early version of the one eventually used for GIVE-2. The three statements about technical problems had "yes/no" options. A 1-7 scale ranging from "never" (1) to "always" (7) was used for 10 items which involved a time dimension. In the remaining items the scale ranged from "strongly disagree" (1) to "strongly agree" (7).

Procedure: After reading a description of the study and signing a consent form, each participant interacted with the GIVE-2 environment while we observed him through the video from another location and took notes. We set a time limit of 10 minutes for the participant to finish the task, after which we would interrupt him and ask him to fill in the questionnaire. Then, one of us interviewed him using the questionnaire as a guide while the other observed through the video and kept notes. The participant was then asked to play the game again and exemplify some of the issues that were flagged during the interview. A debriefing concluded the sessions which lasted between 35 and 50 minutes.

5 Results

5.1 Usability issues

Controls: As we were observing the first participant (S1), we noticed that he placed his left hand on the "w-a-s-d" keys. Instead of navigating, these keys "tilted" the view. This initially

¹www.give-challenge.org/research/page.php?id=software

surprised the participant although he eventually brought the view back to the original angle.

During the interviews, S1 and S3, both experienced gamers, mentioned that in games the “w-a-s-d” keys have the same functionality as the keyboard arrows. This allows gamers to navigate with one hand and use the mouse with the other.

The questionnaire included the statement “I experienced technical problems when using my keyboard to navigate the world” with yes/no as options. Both S1 and S3 chose “no”. Only S5 chose “yes”. In the interview S5 explained that it was hard for him to estimate how long he needed to keep a key pressed in order to move a certain distance towards a particular direction.

Virtual buttons: To press a virtual button the participant has to come close and click on it with the mouse. S1 and S5 were observed a couple of times clicking on buttons without effect because they were not close enough to them. On one occasion, S4 clicked several times on a button without effect before moving closer. All three mentioned the difficulty of pressing buttons in their interviews. We did not anticipate this issue so it did not correspond directly to a questionnaire item.

Identifying objects: All participants said that they had problems identifying the alarm tile on the screen. This was because the tile had the same colour and pattern as the floor. The statement “When the system described objects, it was easy to identify them in the virtual world” was rated with 7 (strongly agree) by one participant and with 5 and 6 by two others.

Poor contrast: Three participants (S2, S4, S5) stated that they could not always read the generated instructions because the contrast between the letters and the background was at times too low. The statement “I experienced technical problems when viewing the world on my screen” was answered with “no” by all three although S4 changed his answer to “yes” during the interview.

Instruction timing: Four participants stated that they could not always read everything on the screen either because the instructions were often not displayed for long enough or because they were coming one after the other in very quick succession. However, S3 mentioned that he could always read the instructions.

Four items in the questionnaire were devoted to timing. The rating of S3 in each item was identical to that of at least one other participant. The

ratings of the four participants who made similar comments varied quite considerably.

5.2 Beyond usability

Prior experience: The online demographic questionnaire which starts a GIVE-2 session includes the question “How many hours per week do you play computer or video games?”. S5 typed “0” since he was not playing games at the time of the study. During the interview, he explained that we used to play much more frequently in the past. This was the case for S3 and S2 too. Only S1 was still playing games frequently.

These four participants found the trophy well within 10 minutes. S1 and S2 were aware of GIVE but had not interacted with the environment before. The statement “With the system’s instructions, finding the trophy was easy” was rated with 7 (strongly agree) by S5 and with 6 by the other three participants.

The inexperienced gamer, S4, (who did not know about GIVE either) had great difficulty navigating the environment and did not find the trophy within the time limit. He remained unaware of the purpose of the game until the interview.

The four experienced gamers said that they understood that they were supposed to look for a trophy soon after they started playing the game even though no precise description was provided to them. However, all participants were uncertain about the function of the alarm tile in the game.

Instead of always relying on instructions, three experienced gamers (S1, S3, S5) found the trophy by following more exploratory strategies. They mentioned that having to wait while the system was updating its plan was “weird” and they simply kept moving. They made limited use of the “H” key which repeats the previous instruction. S5 said that often he did not need to read the whole instruction to figure out what to do next.

Overall perception: S2 said repeatedly that interacting with GIVE-2 felt more like performing a task than playing a game. S3 mentioned that it was less exciting than other games he had played while S5 said that it did not contain enough stimuli and he would not play it again. He also stated that if GIVE-2 were “a real game” he would have tried to “break it” by going through walls, doing the opposite of the instructions, etc. These three participants agreed with the statement “I found the task boring” with ratings of 5 (twice) and 6.

All participants (except for S2) rated the statement “I enjoyed solving the task” positively with 5 and 6 (twice each).² Ratings for the statements “I would like to do the task again” and “I would recommend the task to a friend” were between 4 and 6. We emailed the participants five months later to inquire whether they have played or recommended the game since the study. Three replied (S3, S4 and S5), negatively for both questions.

6 Discussion

6.1 Usability

Soon after the study we sent a brief report to the organisers of the challenge mentioning the issues about the “w-a-s-d” keys, the contrast, the timing, the poor visibility of the alarm tile and the incomprehensibility of its function. Consequently, the tile was made more visible and the contrast between the text and the background was increased. These are salient examples of usability issues that may have impacted on the results of GIVE-2.

One of the NLG engines which later participated in the challenge provided additional information about the alarm. Our study suggests that such clarifications can be helpful to participants.

The difficulties with making precise movement and pressing buttons were spotted during the usability test but required more careful inspection of the recordings and are reported here for the first time. If the challenge relies on the assumption that the participant performs precise actions these may also require more attention in the future.

Observing and interviewing a few participants unveiled several unanticipated issues. The exact nature of these issues would have been harder to capture by the participants’ numerical rankings or their answers to yes/no items in the questionnaire. In fact, their questionnaire responses often glossed over usability problems so additional feedback was required. We also found the comments by the participant who did not complete the task to be as insightful as those by the other participants.

The questionnaire was designed to compare different NLG engines based on a large number of responses. We believe that one should not expect it to also serve as a means of usability testing. Performing formative usability evaluation early in development has become common practice not only in HCI research but also for commercial applications. While this study was limited to just one it-

²This includes S4 who did not actually solve the task.

eration, paying more attention to usability issues may be beneficial in future runs of GIVE to ensure that such issues do not interfere with the outcomes of the challenge.

6.2 Wider issues

According to Koller et al. (2009, p.301), “from the perspective of the users, GIVE consists in playing a 3D game”. Our participants (perhaps with the exception of S2) appeared to enjoy interacting briefly with GIVE-2. However, they stated quite clearly that this felt more like performing a task than playing a game. If gaming is not the targeted domain, it might be worth considering more carefully to which other virtual environments the results of GIVE would apply.

Our study suggests that gaming experience may influence the participants’ ability to finish the task and the strategies that they adopt. That participants are guided by prior experience is not a new finding (see e.g. Weibelzahl and Reynolds (2009, p.117), and Jokinen and Hurtig (2006)). As the example of S5 shows, participants may have considerable gaming experience even though they do not play games frequently any longer. To prevent these variables from masking the results of the challenge, a better understanding of the participating population may be required.

Experienced gamers often did not read the whole instruction and did not always act in response to it. Similar exploratory behaviour in human-machine communication was analysed by Suchman (1987). Her analysis, which has been very influential in HCI, suggests that providing instructions in a stepwise manner based on a hidden plan may impact more on the interaction than the actual linguistic quality of the instructions. This is another issue to consider if the purpose of GIVE is to provide insights about the potential of NLG engines outside that particular environment.

7 Conclusion

Our formative usability test unveiled unanticipated issues which may have otherwise impacted on the results of GIVE-2. Future runs of GIVE may benefit from closer integration of usability evaluation. The study also identified wider issues related to the purpose and the generalisability of the challenge which may be worth more attention in subsequent research.

Acknowledgments

Many thanks to our participants for their time and to Stephan Weibelzahl and Abi Reynolds for their advice and for providing us with access to the usability laboratory of the National College of Ireland. Also many thanks to Kristina Striegnitz, Alexander Koller and Andrew Gargett for their responses to our various questions and to Stephan Schlogl, Anne Schneider, Illana Rosanes and Cecily Morisson for their comments. This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie) at Trinity College Dublin.

References

- D. Byron, A. Koller, K. Striegnitz, J. Cassell, R. Dale, J. Moore, and J. Oberlander. 2009. Report on the First NLG Challenge on Generating Instructions in Virtual Environments (GIVE). In *Proceedings of ENLG 2009*, pages 165–173.
- K. Jokinen and T. Hurtig. 2006. User expectations and real experience on a multimodal interactive system. In *Proceedings of Interspeech 2006*. Paper 1815-Tue2A3O.2.
- A. Koller, K. Striegnitz, D. Byron, J. Cassell, R. Dale, S. Dalzel-Job, J. Oberlander, and Moore J. 2009. Validating the web-based evaluation of NLG systems. In *Proceedings of ACL 2009*, pages 301–304.
- S. Krug. 2006. *Don't make me think!: a common sense approach to web usability*. New Riders.
- J. Nielsen. 1989. Usability engineering at a discount. In *Proceedings of 3rd International conference on Human-Computer Interaction*, pages 394–401.
- L. Suchman. 1987. *Plans and situated actions: the problem of human-machine communication*. Cambridge University Press.
- S. Weibelzahl and A. Reynolds. 2009. Usability Testing of e-Learning in Practice: First Experiences and Lessons Learned. In *Proceedings of Irish HCI 2009*, pages 115–118.