Analysing screen-capture recordings to explore user-experience with Web sites

Gabor Aranyi
Teesside University
Middlesbrough, TS1 3BA
G.Aranyi@tees.ac.uk

Joseph A. Onibokun
Teesside University
Middlesbrough, TS1 3BA
Joseph.Onibokun@yahoo.co.uk

This workshop contribution presents reflections on the collection and analysis of screen-capture video and audio recordings in two separate studies. Participants in both studies used a Web site individually in a laboratory setting under think-aloud instructions, while their online use and verbal behaviour was recorded using screen-capture software. One study used an online news site, the other a social-networking site as interactive artefact. The method of data collection, transcription of protocols, extraction of themes and the analysis of protocols are illustrated by comparing and contrasting the two studies. The use of the presented technique yielded valuable insights into user-reported aspects of user-experience with Web sites, and the findings of the studies were applied in further research to inform the selection of psychometric measures for modelling user-experience to guide system evaluation and design. Examples of results from the studies are provided to demonstrate the usefulness of the applied analysis techniques. Practical implications for collection, transcription and analysis of screen-capture video and audio data are discussed.

Screen-capture *Protocol analysis* *Think-aloud* *User-experience*

1. INTRODUCTION

In recent years, users' experience with interactive technologies has received an increasing amount of attention both in system design and human-computer interaction research. Examples of such interactive technologies include mobile telephones (e.g., Thüring & Mahlke, 2007; Cho et al., 2011), Web applications (e.g., online games; Hsu & Lu, 2003), and retail and educational Web sites (Porat et al., 2007; Hartmann et al., 2008). Research in user-experience aims to understand and describe the source and characteristics of positive experiences with interactive technologies (Hassenzahl et al., 2010; Law & van Schaik, 2010). Promoting high-quality experiences, in turn, is expected to contribute to the success of these technologies. Conducting user-tests and analysing data collected from users are common and important practices both in academic research and industry (Nielsen, 1993). Furthermore, video analysis of user tests can provide rich descriptions of participants' use of interactive products (Vermeeren et al., 2002).

The authors of the present paper were members of a research group at Teesside University, working on separate projects. The methods for the studies were developed and tailored separately by the authors to the needs of the individual projects. Both studies employed video data to identify factors or categories of users' experience, however with different artefacts and theoretical focus. Participants in Study A (reported in Aranyi et al., 2012) used an online news site; participants in Study B (reported in Onibokun & van Schaik, in press) used a social-networking site.

2. METHOD

Both studies were conducted in a laboratory setting with participants using a Web site under think-aloud instructions. Study B recruited experienced users of a particular social-networking site (N = 26), while Study A recruited both users (n = 10) and non-users (n = 15) of a particular news site in order to explore possible differences in reports of novice and expert users. Both studies employed a concurrent think-aloud protocol, while users were instructed to use the sites as they normally would (i.e., there were no set tasks). Researcher-participant interactions during the recording sessions were kept to a minimum in order to facilitate the collection of unbiased protocols (for detailed descriptions of the application of different types of think-aloud protocols, see Ericsson & Simon, 1993; Boren & Ramey, 2000; van den Haak et al., 2004). Both studies were conducted for the purpose of exploring users' self-reported aspects of experience with particular types of Web site. The results were used to inform the selection of psychometric measures for modelling interaction experience with news sites (Study A) and the acceptance of social-networking sites (Study B). Furthermore, results from Study A were used to provide feedback and guidance for the news site's developers.

3. ANALYSIS

Video and audio data in both studies were collected using screen-capture software (Camtasia Studio 3.0). The resulting AVI files contained a recording of the full computer screen during the sessions with the participants' voice, allowing for the integrated analysis of browsing behaviour and think-aloud protocols. Additionally, both studies applied a set of psychometric instruments to measure uses' experience with the Web sites. The two studies are compared and contrasted on the following points.

1. Transcription of protocols
2. Reliability analysis
3. Protocol analysis

3.1 Transcription of protocols

In Study A, each participant (*N* = 25) provided approximately 10 minutes of recording. The recordings were watched repeatedly by the researcher to identify units of thought (Riffe et al., 2005). Units of thought were defined as participants' expression of judgement or opinion about the site, or expressions of their experience. Altogether 190 units were identified. Instead of the full protocols, only the identified units were transcribed for analysis to reduce workload and save time by concentrating on units in the protocol that are relevant to the research aims.

By contrast, the researcher in Study B transcribed the complete protocols as verbatim as possible; a very time-consuming process which, however, increases the reliability of the coding procedure (i.e., every element of the protocol was coded; Young, 2005). Participants in Study B (*N* = 26) were instructed to use the site while thinking aloud for up to 30 minutes. The transcripts were divided into 500 units of thought or comments to serve as the basis of content and thematic analysis.

Screen-capture recordings in both studies were used to identify pages and elements of pages on which the participants' comments were made about a particular property of the site or their experience. The video recordings were also useful to clarify ambiguous elements of verbal protocols, for example, by allowing for the identification of reference for utterances such as "this part of the site" and "that thing in the corner" from the movement of the mouse pointer. Additionally, video information could be used to analyse navigation behaviour in detail; however, this was outside the scope of the presented studies.

3.2. Reliability analysis

Following transcription, researchers in both studies devised categorisation schemes consisting of categories and sub-categories, into which the units of thought were assigned. The development of categories was an iterative process where each unit was treated as a potential member of one or more categories (i.e., the categories were not mutually exclusive). Both studies assessed the reliability of the categorisation scheme by assessing the level of agreement between a researcher and two independent coders using ReCal, an online inter-coder reliability Web service (Freelon, 2010). Fleiss' Kappa, pairwise Cohen's Kappa and Krippendorff's alpha values were assessed using Landis and Koch's (1977) guidelines for degrees of agreement.

Upon devising the categorisation scheme and assigning over 95% of units to the categories, the researcher in Study B randomly selected a sample of 84 out of 500 units of thought (Lacy & Riffe, 1996), and gave the units and a coding sheet to the independent coders to assess inter-coder reliability. Once the reliability of the categories was confirmed, the researcher used his own categorisation solution for further analyses. By using a random sample rather than the full number of units for the reliability analysis, the researcher in Study B managed to significantly decrease the workload for the independent coders.

By contrast, all 190 units of thought in Study A were categorised by three coders. Upon establishing the reliability of the categories, the independent coders' categorisations were incorporated in the analysis: only those units of thought were reported (and quoted) as typical members of a category which the coders agreed upon. Therefore, the independent coders' categorisations were not only used for the assessment of reliability, but were also applied to filter out ambiguous units from the reports, thereby increasing the consistency of the categories.

3.3 Protocol analysis

In Study A, five self-reported categories of experience with a news site were identified from the protocols[1]. Three of the categories were further divided into three sub-categories each to provide a more detailed analysis. In Study B, eight categories of experience with a social-networking site were

[1] The results of the analysis are not described here for reasons of brevity. See Aranyi et al. (2012) for a detailed report of Study A, and Onibokun and van Schaik (in press) for Study B.

identified with no sub-categories. The fit of the categorisation scheme in both studies was adequate: the majority of the units were assigned to at least one of the categories (92% in Study A and 95% in Study B) and satisfactory inter-coder reliability values provided support for the robustness of the categories.

Thematic analysis in both studies included the description of the categories (or themes) with the use of quotes from the transcripts. Motives for using social-networking sites were identified in Study B and were mapped to a selected set of psychological needs (including relatedness, stimulation and popularity; Sheldon et al., 2001), thereby anchoring users' experience in need fulfilment. Study A also assessed users' experience with a news site with a set of psychometric measures, which were then mapped to the categories of thematic analysis.

The categories were described with two values in both studies: size and prevalence. The *size* of a particular category was the number and percentage of all units of thought assigned to the category, indicating the overall frequency of the particular aspect of experience in the protocols. The *prevalence* of a particular category was the number and percentage of all participants who reported at least one unit of thought into the category, indicating the overall prevalence of a particular aspect of experience among the participants. For example, 33% size and 96% prevalence of the category *communication* in Study B meant that 33% of all units regarded an intentional exchange of dialogue between groups and individuals using a social-networking site, and 96% of participants made at least one comment coded as a member of this category[2].

Study A recruited both users and non-users of a particular news site to explore possible differences between novice and expert users. Odds ratios were calculated to each category to express differences between the groups as the number of times the odds of a novice user producing a unit of thought that was assigned to a particular category are greater than that of an expert user (see Field, 2009, Chapter 18). For example, an odds ratio of 2.5 for the *information architecture* category indicated that the odds of novice users producing a unit of thought regarding the navigation, links and structure of the site were 2.5 times the odds of expert users.

The units of thought in Study A were also coded for valence (positive, negative or neutral). Valence analysis showed that the valence of units was not evenly distributed among the categories; for example, negative units outweighed positive units four to one in the *diversion* category, which collated units regarding distraction and confusion in participants, and comments regarding loading time. Furthermore, the examination of odds ratios suggested that although both novice and expert users were likely to make negative comments about the site to a similar extent, expert users were notably more positive in their judgements. These findings support the ideas that level of adoption of a particular artefact may affect users' experiences, and various aspects of experience may involve different valence.

The protocols in Study A were divided into one-minute-long segments and a timeline analysis was carried out, which allowed for the examination (and graphic presentation) of the distribution of units of thought per category from the first to the eleventh minute of recordings. The analysis revealed, for example, that 75% of overall impressions and judgements about the site were expressed in the first minute, while the rest of the categories (e.g., layout and content) showed a more even distribution. Furthermore, the examination of video recordings revealed that 88% of these overall judgements were uttered by participants while on the home page of the site. These findings support the idea that overall impressions form relatively early during use and the home page of a site may be very important in influencing these impressions.

4. SUMMARY

The studies presented in this paper used video and audio data to analyse retrospective think-aloud protocols recorded with users of two different types of Web site. Details of the studies were presented to illustrate a number of issues and considerations regarding collection, transcription and analysis of this type of data. A summary of these considerations is presented in Table 1. Examples of findings from the studies illustrated the usefulness and feasibility of the collection and analysis of screen-capture recordings. Video data in the presented studies were used to guide the analysis of verbal protocols, predominantly by aiding the researcher in the identification of pages and elements of pages as sources of participants' verbalisations regarding their experience, and by the clarification of ambiguous elements in the transcripts. In summary, the analysis of screen-capture video and audio data proved to be a useful technique to explore and identify categories of users' experiences with Web sites. Finally, screen-capture data collection is not limited to the study of Web sites loaded on computers; the method can be used for the study of other screen-based activities.

[2] Note that size and prevalence values are only meaningful as indicators of importance of categories when participants are not prompted as to which aspects of their experience to give account of.

Table 1: *Summary of the studies.*

	Study A	Study B
Artefact	News site	Social-networking site
Participants	Novice/experienced	Experienced
Protocol length	Approx. 10 minutes	Approx. 30 minutes
Transcription	Units of thought only	Verbatim
Reliability analysis	All units of thought	Sample of units
Segmentation of themes	Categories and sub-categories	Categories only
Quantitative analysis of themes	Size and prevalence Between-group diff. Valence analysis Timeline analysis	Size and prevalence
Mapping of themes	Psychometric questionnaires	Need fulfilment

5. ACKNOWLEDGEMENTS

The authors are grateful for Paul van Schaik for his guidance and advice as the research group leader during the design and implementation of the presented studies.

6. REFERENCES

Aranyi, G., Schaik, P. van and Barker, P. (2012) Using think-aloud and psychometrics to explore users' experience with a news Web site. *Interacting with Computers*, 24(2), pp. 69-77.

Boren, M. and Ramey, J. (2000) Thinking aloud: reconciling theory and practice. *IEEE Transactions on professional communication*, 43(3), pp. 292-302.

Cho, Y., Park, J., Han, S. H. and Kang, S. (2011) Development of a web-based survey system for evaluating affective satisfaction. *International Journal of Industrial Ergonomics*, 41(3), pp. 247-254.

Ericsson, K. A. and Simon, H. A. (1993) *Protocol analysis: verbal reports as data*. Cambridge, MA: MIT Press.

Field, A. (2009) *Discovering statistics using SPSS*. 3rd ed. London: Sage.

Freelon, D. G. (2010) ReCal: intercoder reliability calculation as a Web service. *International Journal of Internet Science*, 5(1), pp. 20-33.

Haak, M. J. van den, Jong, M.D.T. de and Schellens, P.J. (2004) Employing think-aloud protocols and constructive interaction to test the usability of online library catalogues: a methodological comparison. *Interacting with Computers*, 16(6), pp. 1153-1170.

Hassenzahl, M., Diefenbach, S. and Göritz, A. (2010) Needs, affect, interactive products - facets of user experience. *Interacting with Computers*, 22(5), pp. 353-362.

Hsu, C-L. and Lu, H-P. (2003) Why do people play on-line games? An extended TAM with social influences and flow experience. *Information & Management*, 41(7), pp. 853-868.

Lacy, S. and Riffe, D. (1996) Sampling error and selecting intercoder reliability samples for nominal content categories. *Journalism & Mass Communication Quarterly*, 73(4), pp. 963-973.

Landis, J. R. and Koch, G. G. (1977) The measurement of observer agreement for categorical data. *Biometrics*, 33(1), pp. 159-174.

Law, E. and Schaik, P. van (2010) Modelling user experience - an agenda for research and practice. *Interacting with Computers*, 22(5), pp. 313-322.

Nielsen, J. (1993) *Usability engineering*. London: AP Professional.

Onibokun, J. A. and Schaik, P. van (in press) Using a classification of psychological experience in social-networking sites as a virtual learning environment. *International Journal of Virtual and Personal Learning Environments*.

Porat, T., Liss R. and Tractinsky, N. (2007) E-stores design: the influence of e-store design and product type on consumers' emotions and attitudes. 12th International Conference on Human-Computer Interaction (HCI) 2007, July 22-27, Beijing. Published in: *Lecture Notes in Computer Science (LNCS): Vol. 4553* (pp. 712-721). Berlin/Heidelberg: Springer.

Riffe, D., Lacy, S. and Fico, F. G. (2005) *Analysing media messages: using quantitative content analysis in research*. 2nd ed. Mahwah, NJ: Lawrence Erlbaum Associates.

Sheldon, K. M., Elliot, A. J., Youngmee, K. and Kasser, T. (2001) What is satisfying about satisfying events? Testing 10 candidate psychological needs. *Journal of Personality and Social Psychology*, 80(2), pp. 325-339.

Thüring, M. and Mahlke, S. (2007) Usability, aesthetics and emotions in human–technology interaction. *International Journal of Psychology*, 42(4), pp. 253–264.

Vermeeren, A. P. O. S., Bouwmeester, K. den, Aasman, J. and Ridder, H. de (2002) DEVAN: A tool for detailed video analysis of user test data. *Behaviour & Information Technology*, 21(6), pp. 403-423.

Young, K. (2005) Direct from the source: the value of 'think-aloud' data in understanding learning. *Journal of Educational Enquiry*, 6(1), pp. 19-33.