Usability Evaluation Issues in Commercial and Research Systems

Laila Dybkjær[#], Niels Ole Bernsen[#], Hans Dybkjær[*]

#NISLab, University of Southern Denmark
Campusvej 55, 5230 Odense M, Denmark
laila@nis.sdu.dk, nob@nis.sdu.dk

*Prolog Development Center A/S (PDC)
H. J. Holst Vej 3C-5C, 2605 Brøndby, Denmark
dybkjaer@pdc.dk

Abstract

This paper briefly reviews current-practice usability evaluation methods and criteria for spoken dialogue systems. We then describe how two commercial and one research system were evaluated with respect to usability and discuss similarities and differences. Finally, we discuss the industrial need for cheaper ways of evaluating usability and the need to pursue research on usability in a field in which the technological capabilities of systems continue to improve and diversify at a rapid pace.

1. Introduction

In recent years, usability evaluation of spoken, possibly multimodal, dialogue systems (SMDSs) has come into focus both as a research topic and as a commercial quality parameter. Usability is likely to remain important as increasing numbers of users with no particular computer skills use electronic devices which can run spoken dialogue applications. In parallel, applications are becoming more sophisticated and powerful, which implies new challenges for usability design and evaluation.

Due to the nature of the technology, industry has never ignored usability entirely. However, focus is still on building the system and ensuring that it has the required functionality. Usability is recognised as important but still seems to belong to the luxury category. Normally, only limited usability evaluation is done unless the customer wants to pay for more. A customer might be willing to pay slightly more for improved usability but only within limits. Extensive usability testing is expensive and may not improve usability with a factor comparable to the increase in price.

We believe that there is reason to address usability evaluation from various angles. *One* is to find ways to get more usability evaluation for less cost. *Another* is to extend our knowledge of which parameters contribute to usability and how much. A *third* is to investigate how to deal with new issues in evaluation imposed by new application types.

This paper provides a brief current-practice overview, presents and compares the evaluation of three very different systems and discusses the above three challenges.

2. Current-practice in usability evaluation

The development of an SMDS is development of a piece of software and hence should follow the software lifecycle process. Over the years, it has become clear that iterative development tightly integrated with evaluation is the most appropriate approach. Evaluation includes technical and usability evaluation. Both should be done from early on and throughout the life-cycle. Before performing usability evaluation one should consider the purpose, the method(s) to use and which evaluation criteria to apply. The purpose influences the choice of methods and criteria. Parameters, such as resources available and the system's stage of development, also have an impact on which method(s) and criteria to use.

2.1. Current-practice usability evaluation methods

No single usability evaluation method can provide answers to everything. Thus it is preferable to use a mix of different methods during development. In the following, we briefly mention a number of current-practice usability evaluation methods with recommendations for when to use them and comments on their drawbacks and advantages. The list is not exhaustive. See, e.g., [19] for a broader discussion.

Some methods are primarily meant for early evaluation with no implemented system, e.g., walkthroughs, mockups or paper prototyping; some aim at the partially implemented system, e.g., high-fidelity prototyping, "bionic" Wizard of Oz, and controlled laboratory testing; others are for the phase when the system is fairly close to completion, e.g., controlled laboratory testing and field tests; yet others are applicable at any time during development. Interviews, questionnaires and think-aloud protocols may be used throughout and in combination with any method involving users. Heuristic evaluation and expert reviews may be used any time but preferably in the early and middle phases to supplement a method involving representative users. The earlier errors can be caught and corrected, the cheaper the development process will be.

Early methods typically involve a system version which is easy to set up and change. A drawback is that such versions are far from being a real system. Yet these methods can generate useful insight and reveal major usability problems.

Methods involving users are generally costly as they involve (i) efforts to find users, prepare scenarios, and make sure everything works, and (ii) analysis of the data collected during interaction and possibly afterwards in interviews or questionnaires. The analysis process is typically time consuming. Still, it is crucial to get users' reactions to the system to detect major inadequacies early on. Care should be taken that the collected data is reliable and not, e.g., corrupted by the use of priming scenarios or leading questions.

Heuristic evaluation and even expert reviews are cheap. Heuristic evaluation requires a set of guidelines and the sole focus is on whether the system follows the guidelines. An expert review requires that an expert can be found and findings may be tainted by the opinions of the expert.

In addition to cost and development phase the choice of method depends on evaluation purpose. Overall, we may distinguish between diagnostic, performance and adequacy evaluation. Diagnostic evaluation finds and diagnoses errors to help repair the system. Performance evaluation measures user performance with the system. Adequacy evaluation concerns how well the system fits its purpose and meets user needs and expectations. Thus, e.g., expert reviews and heuristic evaluation may work well in diagnostics but are unsuited if

focus is on user performance. For realistic measurement of many aspects of performance, an implemented system is needed whereas adequacy may be measured using, e.g., a Wizard-of-Oz simulated system.

A fair range of evaluation methods exist which can help developers improve usability no matter if the system is a research system or a commercial one. The main problem in the use of those methods is probably one of cost, in people, time and money. There must be people with the skills to select and apply suitable methods, time to conduct usability testing and analyse results, and a reasonable budget for doing it.

2.2. Current-practice usability evaluation criteria

Evaluation is quantitative or qualitative, subjective or objective [6]. Ideally, we would like to have quantitative and objective usability evaluation scores which, e.g., can be objectively compared to scores obtained from other SMDSs or previous versions of the same SMDS. Currently, many important usability parameters cannot be quantified, and objective expert evaluation can be uncertain or non-existent. Thus, subjective evaluation remains crucial in usability evaluation.

The ISO standards for measuring software usability and quality can be used for SMDSs evaluation. ISO 9241-11 on usability lists effectiveness, efficiency and satisfaction but many other parameters are relevant. Thus, at [12], usability is also defined as having to do with (i) learnability, including predictability, synthesisability, familiarity, generalisability, and consistency, (ii) flexibility, including dialogue initiative, multi-threading, task migratability, substitutivity, and customisability, and (iii) robustness, including observability, recoverability, responsiveness, and task conformance. ISO/IEC 9126-1 on product quality [13] introduces parameters, such as attractiveness, understandability, and operability.

The SMDS literature proposes many other criteria, e.g., modality appropriateness; adequacy of modality understanding, output phrasing, output representation, error handling, feedback, and emotion expression; quality of output, e.g., voice, graphics, and animation; naturalness of interaction and embodied agent; ease of use of system and devices; frequency of interaction problems; sufficiency of domain coverage, reasoning capabilities, and user modelling; task success rate, error correction rate, and marketability; see also the list of quality features in [16].

A major problem, therefore, is to select the right criteria for the test of a given system. For an evaluator it is important to know the range of criteria available, to make a proper selection for the purpose. A second major problem is that many usability criteria are vaguely defined, making them hard to apply. For example, when is it safe to conclude that system X is “adequate” in respect Y? New system types may require new criteria to be clearly defined and operationalised.

3. Usability evaluation of three systems

While the same set of methods and criteria is available to industry and academia, their actual use differs across systems, as illustrated for the systems described in the following, i.e., a traffic information system, a frequently asked questions system, and an edutainment system. Part of the difference can be explained by differences in system complexity. Increased complexity seems likely to require an increased effort in usability evaluation. A second factor is that usability evaluation cost must be kept low for industrial systems to be competitive wrt. price, whereas it may be a research aim to study usability evaluation aspects and therefore put considerable effort into evaluation. For further information on evaluation activities see [7], and see [16] for a brief overview of how 15 different systems were evaluated.

3.1. Traffic information system

The traffic information system is a telephone-based commercial dialogue system which was put in operation in spring 2005. It informs about overall traffic conditions in major regions of Denmark and in particular about traffic and delays due to construction work on a motorway in Copenhagen. The system, developed by PDC, is not very complex and clearly feasible within the limits of today’s SMDS technology.

Dialogue model design and evaluation was supported by a tool, DialogDesigner [5], cf. spokendialogue.dk. DialogDesigner includes a Wizard of Oz tool which was used for developer walkthroughs of the dialogue model and semiformal Wizard of Oz sessions with colleagues. Focus was on pin-pointing interaction problems and missing functionality, and checking for correctness. After implementation, 8 colleagues who were not in the project were given scenarios and called the system. Each had three scenarios one of which left it completely to the test person what to ask for. Four scenario sets were used, totalling 8 scenarios plus the free one. Again, focus was on possible interaction problems and functionality. The dialogues were not analysed in detail, neither with the simulation nor with the implemented system. Identification of problems was based on observations during the interactions and feedback from the test subjects. Logs were used for the problems analysis.

3.2. Frequently asked questions system

The FAQ system is a telephone-based commercial dialogue system which was put in operation in 2002. The system provides general information on holiday allowance and answers questions, such as “Is Saturday considered a holiday” or “Can I transfer vacation to next year”. Development was supported by a government grant to stimulate uptake of spoken dialogue systems in Denmark, the FAQ system being borderline of commercial feasibility due to its large unstructured domain. PDC and NISLab collaborated on its development.

A two-step approach was used. First, a limited FAQ called Vejled (Guidance) was developed to get the technology in place while still having a relatively simple dialogue, and to generate initial experience with real users. In particular, the initial prompt design turned out to be crucial. The second step was to enhance Vejled into a real FAQ system.

The first tests of Vejled were internal with colleagues. Focus was on identifying missing functionality and usability problems during interaction. All calls were transcribed and analysed, resulting in changes to the system. The amount of data was small as people didn’t phone more than once unless explicitly told. In spring 2002 we invited people outside our sites to call Vejled. This resulted in 225 calls which were transcribed. Dialogue transactions were carefully analysed since transaction success was a key metrics in the contract. People were encouraged to fill in a questionnaire but only 12 did. Users were quite positive on average but clearly there was still room for improvements. Vejled was improved and put in production in summer 2002. All collected dialogues

were transcribed and some were selected for further analysis to provide input to the FAQ system.

The FAQ system was also initially tested internally and rather systematically to see if all required functionality was in place and if there were problems in the interaction. Later we made closely monitored lab tests with external people, each subject carrying out a number of scenarios. After interaction, the test person was interviewed to get his/her opinion on the system and any encountered problems. All FAQ dialogues were transcribed and, among other things, analysed for interaction problems. After the FAQ was put into production we continued to receive a batch of 150-1000 dialogues every week, depending on the season, all of which were transcribed during the first five months of 2003. About 300 of these were selected for analysis and annotated for transaction success.

3.3. Edutainment system

The Hans Christian Andersen (HCA) system was developed in EU research project NICE (2002-2005) [18]. The system enables spoken and 2D gesture interaction with 3D embodied conversational fairytale author HCA in his study. Natural language understanding, conversation management and response generation was developed by NISLab. NISLab conducted closely similar user tests with the first and second prototypes (PT1 and PT2). Representative target users (children aged 10 to 18 years) were used in both tests. The users interacted with HCA in two sessions of 15-20 minutes. In Session1 they were entirely on their own regarding what to talk to HCA about. In Session2 they used a sheet with suggestions for what they might try to talk to him about, such as finding out about his family or the pictures in his study. After interaction, each user had a structured interview. PT2 gave rise to more questions than PT1, due to added functionality. The PT1 and PT2 tests are reported in detail in [2][3].

Decisions on which usability issues to evaluate were made early in the project. Basic usability evaluation criteria included: adequacy of speech and/or gesture understanding; quality of output voice, animation, and graphics; adequacy of output phrasing; sufficiency of domain coverage; number of objects interacted with; number of topics addressed; and frequency of interaction problems. Core criteria included: conversation success; naturalness of using speech and gesture and of output behaviour; sufficiency of reasoning capabilities; ease of use; adequacy of error handling; scope of user modelling; entertainment value; educational value; and user satisfaction. Many of these criteria required subjective evaluation which is why the interviews were crucial.

3.4. Comparison of evaluation processes

Comparison among the evaluation processes just described suggests that the larger and more unstructured the domain, the more complex the system, and the more research involved, the larger is the need to repeatedly collect and analyse test data. For the research system, a larger number of criteria were used than for the commercial ones. It is natural that innovative SMDS research requires application of a broad range of criteria for usability evaluation.

For all three systems, usability evaluation is heavily based on empirical methods for the study of user-system interaction. Interaction problem analysis is considered crucial in all cases. Interaction problems are closely related to measures, such as task/conversation success, ease of use, and error handling adequacy. Moreover, domain coverage is considered an important usability parameter for all three systems.

User satisfaction is also important. User satisfaction has not yet been investigated for the commercial systems but this will be done later for the traffic information system. During development, focus was not on obtaining an artificial user satisfaction rate as it would have been, had we asked our subjects. Focus was rather on eliciting input on what should be improved and how they had experienced the interaction. Some overall satisfaction rate might have been deduced from their answers to questionnaires and interviews but we did not do that. For the research system, users were asked a number of interview questions which could indicate their degree of satisfaction, such as their general evaluation of the system.

Heuristic evaluation was not made of any of the systems. Expert evaluation was made of the traffic information and holiday allowance rules by the customers' domain experts. For the FAQ system, an expert review was made by NISLab as the non-developing site. In the NICE project, no expert review was made. Indeed, it would have been difficult to find an expert who had the time and skills required for this purpose.

4. Industrial needs and challenges

To reduce usability evaluation cost we seem to need either to automate parts of the process which today are done manually or find new low-cost evaluation methods. The ideas of simplifying or automating evaluation are not new. Several evaluation frameworks have been proposed in recent years, of which PARADISE [21] is probably the most well-known. The idea is to predict user satisfaction from quantitative metrics, such as elapsed time and number of turns, avoiding reliance on subjective user opinions from usability tests. Various weaknesses have been pointed out, however [7][15][16], and prediction accuracy is not good enough [10][17]. Building on PARADISE, the PROMISE framework [1] was applied to multimodal interaction in the SmartKom project and later modified [20] but there are still several open issues.

(Partial) automation has the potential to effectively reduce evaluation effort. The following examples address work towards automation. If interaction involves speech recognition, there are tools which facilitate transcription by displaying what was recognised. It is much faster to correct misrecognitions than typing everything from scratch.

Transaction or task success rate is often considered an important measure and work has been done to automate the process of annotating transaction successes, including the problems involved in achieving success. Hastie et al. [11] infer task completion from tagging of specific system utterance states. They disregard user utterances although, in principle, task completion needs not equal task success. Dybkjær and Dybkjær [4] investigate the possibility of automatic derivation of transaction success for task-oriented dialogues from simple act-topic annotations that also provide an idea of dialogue smoothness.

Harris [9] stresses the importance of an electronic dialogue model. Once we have an electronic model we can add various kinds of support for development and evaluation. The DialogDesigner tool is an example and there are others, see, e.g., [14][22]. DialogDesigner supports the design of a dialogue model, including graphical presentation, for relatively simple, task-oriented dialogue. This model can then

be used for Wizard of Oz simulation and for generation of test scripts for later evaluation of the implemented system. Other evaluation support might be added to DialogDesigner, e.g., transcription support, a check for blind ends and coherence in the dialogue model, cf. [22], or well-formedness checks according to, e.g., some kind of act-topic pattern rules.

5. Research project challenges

SMDSs still have a long way to go before we can emulate multimodal human natural interaction abilities. On the way towards this goal, when pushing the limits of current technology, new usability challenges are bound to arise continuously in the form of new factors to analyse, define, and measure, such as, right now, conversation success, educational value, entertainment value, or the roles and uses of animated interface agents.

We still do not know exactly what user satisfaction is or how it can be reliably predicted from en ever-expanding set of relevant usability parameters. Among other things, user satisfaction is a function of users' preferences, and these differ from one user to another. This emphasises the importance of research on on-line adaptive systems which can also adapt to differences in user skills and background knowledge, and which can do so quickly, given the fact that many SMDSs are meant for walk-up-and-use. Long-term studies of usability is another important issue. There are not yet many results that show what happens to regular system users' perception of usability over time. Also, it is a moot question today if new modality combinations might help remove some of the familiar SMDS usability issues for good.

6. Conclusion

We have briefly outlined current practice in usability evaluation methods and criteria for SMDSs, followed by comparison of the usability evaluation made of two commercial and one research SMDS of different complexity. On this background, we discussed industrial needs for, and research challenges in, usability evaluation. Important issues to address were found to include increased automatic support for evaluation, research on system and user adaptation, long-term studies of usability, and improved operational definitions of important new, and even some old, evaluation criteria.

7. References

[1] Beringer, N., Kartal, U., Louka, K., Schiel, F., and Türk, U.: PROMISE - A Procedure for Multimodal Interactive System Evaluation. Proceedings of the LREC Workshop on Multimodal Resources and Multimodal Systems Evaluation, Las Palmas, 2002, 77-80.

[2] Bernsen, N.O. and Dybkjær, L.: Evaluation of Spoken Multimodal Conversation. Proceedings of ICMI 2004, Penn State University, USA, 2004, 38-45.

[3] Bernsen, N.O. and Dybkjær, L.: User Evaluation of Conversational Agent H. C. Andersen. Proceedings of Eurospeech, Lisbon, Portugal, 2005.

[4] Dybkjær, H. and Dybkjær, L.: From Acts and Topics to Transactions and Dialogue Smoothness. Proceedings LREC'2004, Vol. V, Lisbon, Portugal, 2004, 1691-1694.

[5] Dybkjær, H. and Dybkjær, L.: DialogDesigner – A Tool for Rapid System Design and Evaluation. In [8], 2005, 227-231.

[6] Dybkjær, L. and Bernsen, N.O.: Usability Issues in Spoken Language Dialogue Systems. In Natural Language Engineering, Special Issue on Best Practice in Spoken Language Dialogue System Engineering, Volume 6 Parts 3 & 4 September 2000, 243-272.

[7] Dybkjær, L., Bernsen, N.O. and Minker, W.: Evaluation and Usability of Multimodal Spoken Language Dialogue Systems. Speech Communication, Vol. 43/1-2, Elsevier, 2004, 33-54.

[8] Dybkjær L. and Minker W. (Eds.): Proceedings of the 6st SIGdial Workshop on Discourse and Dialogue. Lisbon, Portugal, 2-3 September, 2005.

[9] Harris, R.: Voice Interaction Design. Morgan Kaufmann Publishers, 2005.

[10] Hassel, L. and Hagen, E.: Evaluation of a Dialogue System in an Automotive Environment. In [8], 2005, 156-165.

[11] Hastie, H. W., Prasad, R. and Walker, M.: Automatic evaluation: Using a DATE Dialogue Act Tagger for User Satisfaction and Task Completion Prediction. Proceedings of LREC'2002, 641-648.

[12] ISO-defs.1: http://www.tau-web.de/hci/space/i7.html

[13] ISO-defs.2: http://www.hostserver150.com/usabilit/tools/r_international.htm

[14] Klemmer, S. R. Sinha, A. K., Chen, J., Landay, J. A., Aboobaker, N. and Wang, A.: SUEDE: A Wizard of Oz Prototyping Tool for Speech User Interfaces. CHI Letters, The 13th Annual ACM Symposium on User Interface Software and Technology: UIST 2000. 2(2): 1-10.

[15] Larsen, L. B.: Assessment of Spoken Dialogue System Usability – What are We really Measuring? Proceedings of Eurospeech 2003, 1945-1948.

[16] Möller, S.: Quality of Telephone-Based Spoken Dialogue Systems. Springer, USA, 2005a.

[17] Möller, S.: Parameters for Quantifying the Interaction with Spoken Dialogue Telephone Services. In [8], 2005b, 166-177.

[18] NICE project: www.niceproject.com

[19] Preece, J., Rogers, Y. and Sharp, H.: Interaction Design: Beyond Human-Computer Interaction. John Wiley & Sons, 2002.

[20] Rieser, V., Kruijff-Korbayová, I. and Lemon, O.: A Corpus Collection and Annotation Framework for Learning Multimodal Clarification Strategies. In [8], 2005, 97-106.

[21] Walker, M., Litman, D., Kamm, C. and Abella, A.: PARADISE: A Framework for Evaluating Spoken Dialogue Agents. Proceedings of ACL, 1997, 271-280.

[22] WOZ tool: http://www.softdoc.de/body/home.htm