

THUS SPOKE THE USER TO THE WIZARD

Hannes Pirker, Georg Loderer⁺, Harald Trost⁺
{hannes,harald}@ai.univie.ac.at

Austrian Research Institute for Artificial Intelligence (ÖFAI) Schotteng. 3, A-1010 Vienna, Austria *

⁺Department of Medical Cybernetics and Artificial Intelligence University of Vienna
Freyung 6, A-1010 Vienna, Austria

Abstract

Wizard-of-Oz (WOZ) simulations are a popular means for investigating the properties of human-computer interaction. In this paper the findings from a WOZ experiment for evaluating different design options for a spoken dialogue system are presented.

In addition to the documentation of the outcomes of this evaluation in terms of standard quantitative measures we also present findings from a more qualitative analysis of the speech data collected throughout this experiment. It is argued that such a combined analysis of all aspects of the human-computer interaction allows for a correct interpretation of the results and their fruitful application in the context of system prototyping.

1 INTRODUCTION

Dialogues in human computer interaction may vary significantly from dialogues in natural conversations. Wizard-of-Oz (WOZ) simulations are a popular framework for investigating peculiarities of this interaction in general and for the development and evaluation of designs for spoken dialog systems in particular [2].

The WOZ experiment presented in this paper was designed for the purpose of examining different possible strategies for initiative, system feedback, and confirmation requests in a hypothetical dialog system.

Nevertheless this paper is not restricted to the documentation of the effects of different system designs on standard *quantitative measures* – such as number of turns or elapsed time – and satisfaction ratings. Also results gained from *qualitative observations* of the subject’s behaviour such as the gradual adaption of speech style are discussed. This contributes to a better understanding of man machine interaction and also sheds some light on the peculiarities of the WOZ paradigm.

2 VARIATIONS IN THE DIALOGUE DESIGN

For the WOZ experiment a cinema ticket reservation system was simulated [4]. For a successful reservation the

*This work has been sponsored by the *Fonds zur Förderung der wissenschaftlichen Forschung (FWF)*, Grant No. P13224. Financial support for ÖFAI is provided by the Austrian Federal Ministry of Science and Transport.

information on the name of the requested film, its starting time, and the number and location of seats had to be contributed by the user.

This section presents the basic parameters used for the implementation of the system variants.

- **Dialogue initiative**

System initiative or mixed initiative is used, i.e., the WOZ either asked “*Which film do you want to attend?*” or “*How can I help you?*”¹. The latter mode was only used for the initial turn, though.

- **Confirmation**

Confirmation information is offered either directly (“*You want to see the film Titanic?*”²) or indirectly (“*When do you want to see the film Titanic?*”). Confirmation can either be requested by the system after each user turn or in a summarized form later in the dialogue (“*I am repeating your reservation data: ...*”).

- **Information about options**

In case tickets for particular requested seats are not available the WOZ can either simply reject this reservation (“*Sorry. This seat is occupied. Where do you want to sit?*”), give information on available seats (“*In this row there are only free seats at the sides.*”) or offer them actively (“*Do you want to take these places?*”).

- **Recognition performance**

Although the system’s speech understanding component was completely simulated by the WOZ no perfect recognition was pretended. Recognition errors were systematically inserted throughout the dialogue. Additionally, all utterances that included hesitations, repairs, nonlexical fillers (e.g., *uh*), sloppy articulation etc. were immediately penalized by a “*Sorry, I did not understand you. Please repeat your input.*”.

Barge ins were completely ignored.

3 EXPERIMENTAL DESIGN

For the experiment three variants of the ticket reservation system were simulated. For all systems the technical procedure was identical. Subjects were asked to participate

¹The whole experiment was performed in German. For convenience of reading only English translations are quoted.

²In German this word order can be used for both questions and statements.

at the evaluation of an experimental automatic reservation system. They were placed in front of a computer terminal in order to enforce the impression of working with a real computer based system and their spoken interactions were transmitted to the WOZ who was sitting in another room. All the Wizard's utterances were produced by a speech synthesizer. For this purpose the WOZ used a graphical user interface that included facilities for the playback of predefined tokens as well as templates that could be edited at run time. This layout minimized typing requirements for the WOZ and maximized the uniformity of the system's responses. All utterances were synthesized with a slowly declining completely flat pitch contour at a slow speech rate.

The following three reservation systems were simulated by the WOZ:

- **System S1:**

- *System initiative* is used exclusively.
- *Confirmation* is only requested in summary after all the input information has been gathered.
- The layout of S1 thus resembles a touch-tone based system that keeps the initiative.

- **System S2:**

- *Mixed initiative* is implemented to a modest degree, i.e., the system starts with a general "How can I help you?" but returns to system initiative afterwards.
- *Explicit confirmation* is requested after each input.

- **System S3:**

- *System initiative* is used again.
- An *implicit confirmation* strategy is implemented.
- Only this system offered *information on free seats* in case requested seats were occupied already.

A total of 26 subjects (19 male, 7 female) in the age from 18 to 40 and diverse educational and professional background voluntarily participated at the experiment. This population was divided into three groups, one for each system version, i.e., every subject only communicated with a single system.

Each subject performed the following three reservation tasks of increasing complexity.

- **Task T1:** Film: Titanic, Time: 19:00. Reserve 2 tickets in the last row/midth.
- **Task T2:** ... Reserve 2 tickets in the last row/midth. In case these seats are occupied try to reserve another seat in the last row. Otherwise apply the same preferences to the penultimate row etc.
- **Task T3:** ... Reserve 2 tickets. Accept seats in the middle of one of the last three rows only. Otherwise try to switch to a later show.

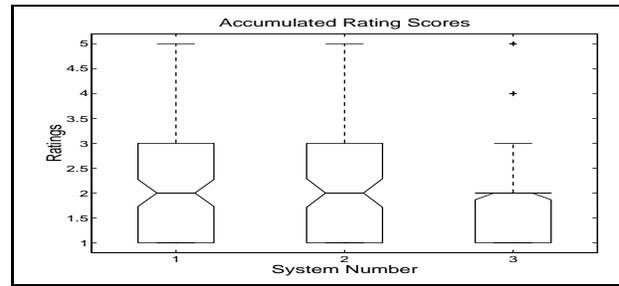


Figure 1: Boxplot of joined ratings for the three systems

After the experiment all subjects answered a questionnaire that contained rating scales as well as open questions about their impression of the system. Also an additional informal interview was performed with about half of the test subjects as well.

A total of 77 evaluable dialogues were produced which comprise a total of 970 turns. The summed transaction time was approximately 3 hours.

4 QUESTIONNAIRE ANALYSIS

The questionnaire used included 20 questions with a five class ranking scale between 1 (agreement) and 5 (disagreement). These comprised questions on the overall performance like

- **Humanpref** I prefer a human operator
- **Util** The system is helpful as it is
- **Easytalk** The system is easy to talk to
- **Easylist** The system is easy to understand
- **Responsespeed** The system responds fast enough

as well as more detailed questions on the design of the dialogue, the easiness of the overall task etc.

Fig.1 presents a plot³ of the accumulated ratings for the three systems. In the accumulated display all factors except **Humanpref** and three control factors with reversed ranking scale had been combined. The median uniquely lies at 2, i.e. in the positive range, for all three systems. System three seems to outperform the other two in this ranking.

A non parametric statistical test was applied in order to check whether all three systems perform identical (null hypothesis). This hypothesis was *not* rejected, i.e. no overall effect of the system was confirmed (type 2 error = 0.05).

In Fig.2 the ratings of some of the global factors were accumulated for all three systems. The high ranking for **Humanpref** (unsurprisingly) indicates the superiority of human systems. Nevertheless the **Util** ranking concedes the system usefulness to some extent.

³The BOXPLOT boxes used in these figures have lines at the lower quartile, median, and upper quartile values. The lines extending from each end of the box (whiskers) show the extent of the rest of the data while outliers are marked with "+". The notches represent an estimate of the uncertainty about the means.

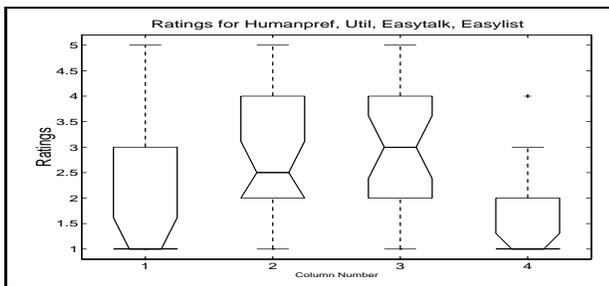


Figure 2: Ratings of global system parameters accumulated for all systems (from left to right: **Humanpref, Util, Easytalk, Easylist**)

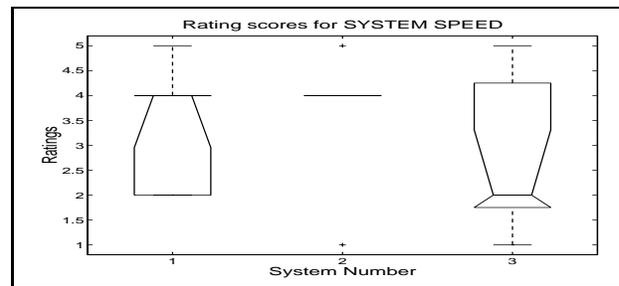


Figure 4: Comparison of the subject’s ratings of **Responsespeed** for the three systems S1, S2, S3.

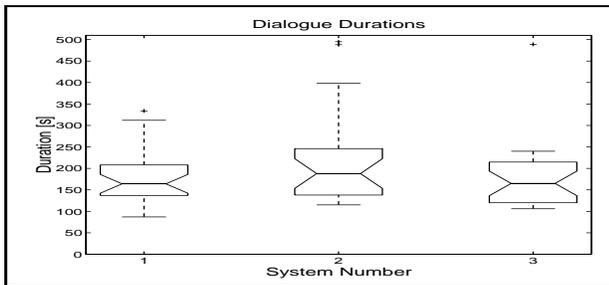


Figure 3: Elapsed time per dialogue for S1, S2, and S3

In the simulation 137 of the 970 user turns are at once explicitly rejected by the WOZ, not counting instances of wrong recognitions. This bad recognition rate is reflected to some extent in the **Easytalk** factor while the rating of the synthesizer’s intelligibility **Easylist** is remarkable high.

5 QUANTITATIVE ANALYSIS

The dialogues were orthographically transcribed. The transcription was labelled for turns and utterances. Overall elapsed time per dialogue as well as number of turns were evaluated. Task success was not included as it was 100% in all systems.

Comparing the overall transaction time in Fig.3 the median of the three systems lies in the range of 2:40 min but there are instances found in both S2 and S3 where transactions lasted up to 8 minutes.

The distribution of the number of turns is similar to that of the durations.

6 OBSERVING USER BEHAVIOUR

As mentioned in the introduction we are convinced that the virtues of WOZ experiments extend beyond their use as a statistical test for the evaluation of system variants.

The analysis of the transcriptions, the sound recordings, and the interviews reveals a whole wealth of possibly unsystematic and anecdotal but nevertheless valid and valuable insights on human behaviour within such a setting.

6.1 Aiding Interpretation

Given the restricted size of the set of available samples in the statistical analysis the role of the qualitative analysis for possibly reevaluating statistical findings becomes even more evident.

To give an example. For all systems the response speed of the WOZ was equally slow due to long pauses and very slow speech rate – too slow in the opinion of the authors. Nevertheless Fig.4 displays different **Responsespeed** ratings for the systems. All the ratings display the general tendency to eschew harsh rejections (rating 5) though. Note that system S2 gets the worst ratings for **Responsespeed** and at the same time displays the longest durations in Fig.3. Nevertheless the correct interpretation of this finding is not straightforward. On the one hand S2’s users might get more annoyed due to their overall longer exposure to the slow system. On the other hand both longer durations and sharper rejection might be due to the fact that by accident a bigger portion of the S2 group seemed to have a more technical background and thus tried to test the system’s abilities and had been more pointed in their overall ratings. Also complaints about the system’s speed were more frankly addressed within the informal interviews than in the rankings.

The rating of **Easylist** in Fig.2 is another example for the importance of performing informal inquiries. While the ratings for **Easylist** are rather positive the interviews revealed that most users were somewhat irritated by the rather low pitched male voice of the system as they awaited and preferred a female voice for the reservation task.

6.2 Speech style adaption

Another strand of observations to be mentioned deals with the expression of attitudes towards the system used. One interesting effect is the adaption of the speaker’s behaviour to the expected language abilities of the counterpart. This effect is well attested for both natural and human-machine dialogues [6]. In the context of our WOZ experiment the subject’s verbal behaviour reveals some information about their conception of the system.

For instance, 8 of the 26 subjects use the politeness marker “please” in their dialogue, typically combined with an overall natural speech style (reflected in articulation and prosody). In the course of the dialogue this speech mode usually is abandoned in favour of slower and more articulated speech and in case of repetitions requested by the system the “please” is omitted. Nevertheless there seems to be no general training effect towards

a more “denaturalized” speech style. This is manifested in the fact that the majority of these “polite” users reapply both natural articulation and politeness markers in more than one dialogue.

6.3 Attitudes

Though most subjects behaved remarkable cooperative with the system (all dialogues where successfully finished at last) low speed and recognition rate frequently provoked emotional attitudes such as mock or impatience which could be traced in both speech and nonspeech behaviour. For instance sighs and loud exhaling was frequently recorded throughout slow and longish system responses.

Some subjects also clearly displayed a switching between “on and off records” speech mode. While the system was answering they frequently commented on erroneous performance, using a casual speech style and directly addressed the system in a “personalized” manner. Thus side remarks like “*Oh. Now you got a problem*” certainly would be rather strange in a natural reservation task. On the other hand in natural conversations subjects are usually not forced to stubbornly repeat the same words and phrases over and over again and have to wait for a response in the meantime.

A prosodic phenomenon was the switching to a sort of stylized intonation contour – sometimes resembling a “calling contour” – which lends the utterances a certain ironic undertone.

6.4 Other issues

In [6] it was demonstrated that users of natural language based systems are influenced by the system’s language. They tend to adjust their utterances in respect to vocabulary and length of phrases. Thus the outcomes of the WOZ experiment inevitably will be biased to some extent by the design of the WOZ which influences the user’s conception of the system’s capabilities. This was also reflected in our experiment where subjects complained in the interview that it was not possible to pose questions to the system, e.g., in order to ask which seats are still free. Nevertheless these subjects did not even try to ask. Because of the system’s poor recognition rate and the predominance of system initiative they automatically assumed that questions were not supported.

7 CONCLUDING DISCUSSION

One of the main experiences gained throughout the WOZ experiment described in this paper is the insight that results from WOZ simulations have at first hand be considered what they are: namely laboratory results. In every single case it has to be carefully evaluated whether these results can readily be generalized to real applications.

Of course this especially applies to the interpretation of subject’s ratings of quality measures which may be used for the comparison of different versions of a system but do not give any reliable information on the “absolute” quality of a dialogue system.

Recent efforts on the development of testbeds for spoken dialog systems and the automatization of the evaluation procedure (e.g., [5],[3]) help to narrow the set of uncertainties in the evaluation procedure by increasing the

grade of formalization. We nevertheless still have to be aware of possible fallacies.

In our experimental setting where all subjects were volunteers and most subjects were not familiar with natural language systems at all, two effects could be detected. On the one hand most subjects seemed to apply a “*principle of mercy*” when completing their rating scales. At the other hand subjects reported to be afraid of “behaving not smart enough” throughout the simulation, i.e., they felt as *they* were tessees instead of testers.

Last but not least in all experimental settings of that sort subjects are basically role playing. They are no real users with real information requirements, real time constraints or real telephone bills. We thus should be careful in our evaluation of results and use them as valuable means for the evaluation of system internal design options but be careful about the generalization of findings.

REFERENCES

- [1] Alter K., Buchberger E., Matiasek J., Niklfeld G., Trost H.: VIECTOS: The Vienna Concept-to-Speech System, in Gibbon D.(ed.), Natural Language Processing and Speech Technology, Mouton de Gruyter, Berlin, 1996.
- [2] Dahlbäck N., Jönsson A., Ahrenberg L.: Wizard of Oz studies - why and how, Knowledge Based Systems, Vol. 6/4 pp.258-266, 1994.
- [3] Litman D.J., Pan S.: Empirically Evaluating an Adaptable Spoken Dialogue System, Proc. of 7th International Conference on User Modeling (UM’99), Banff, Canada, 1999.
- [4] Loderer G.: Evaluierung von Dialogstrategien eines natürlichsprachigen Dialogsystems durch Wizard-of-Oz Experimente (Evaluation of dialogue-strategies in a natural language system using Wizard-of-Oz experiments), Institut für Med.Kybernetik und AI, Universität Wien, Masters thesis, 1998.
- [5] Walker M.A., Fromer J.C., Narayanan S.: Learning Optimal Dialogue Strategies: A Case Study of a Spoken Dialogue Agent for Email, in Proc. of COLING/ACL 98, Université de Montreal, Canada, pp.1345-1351, 1998.
- [6] Zoltan-Ford E.: How to Get People to Say and Type What Computers Can Understand, International Journal of Man-Machine Studies, 34, 527-547, 1991.