Simulation Experiments and Prototyping of User Interfaces in a Multimedia Environment of an Information System

Ludwig Hitzenberger & Huberta Kritzenberger

Dept. Linguistic Information Science
University Regensburg

ABSTRACT

This paper describes some experiments on multimedia user interfaces within the domain of information systems. The factors tested are the input mode (spoken and written) and system quality (cooperativity, language restrictions) and the influence of the consciousness about the interlocutor's identity. The results show that there are some significant differences on the word level (type, token) and that the users reactions on machine performance are more obvious than on the interlocutor's identity.

1 INTRODUCTION

There are good reasons to assume that people talk differently when talking to machines than when talking to other people. They have an imagination of what is the best way of man–machine-conversation, which is not necessarily the most convenient way of processing for the machine. Moreover, machines offer a variety of possibilities of input/output channels such as spoken language and graphic or written screen displays. This makes the proper design of user interfaces cumbersome, since our understanding of these processes is still very poor. The best way would be to test systems in real life applications, where unfortunately most of design decisions are hardly revisable.

Empirical investigations on this subject usually concentrate on the point of the performance of systems using various input/output channels. However speech technology fails with more complicated tasks with complex usage of language. Also, these experiments say nothing about the dependencies of these channels in a multimedia environment, i.e. what is the best use for what facility.

Experimental investigations therefore are complicated, because of the many factors, which are hardly to be isolated. Falzon et al. 1988 report of an experiment, where the only factor was the interlocutor's identity (man or machine). The results, that users complexify their language when communicating with a machine are contradictory to other experiments, also to ours. The reason may be, that intervening factor as the use of the telephone as sole communication channel, the fact that the users did not face the machine or the unnatural way of information display changed the results. We could observe similar effects. Our experimental design therefore stresses the machine environment and the appropriate presentation of the information as a cue factor.

2 METHOD

2.1 TEST DESIGN AND GENERAL CONSIDERATIONS

In a hidden operator experiment we simulate three different information systems, since the complexity of the application domain has a strong influence on the interaction. The domains are a railway information system, a library information system and an office environment. The layout of the system is as close to reality as possible using real data if possible. The information provided for the user is given in the most appropriate way either on CRT in a graphic or textual presentation or with other media if it appears useful. The amount of information given, we derived from a study of real life dialogues from the respective information desks. Our interest is the influence of machines on human behavior, so the categories we investigate are features of computertalk, i.e. deviant language behavior (cf. Capellmann/Franzke/Krause 1988). Besides the common counting of types and tokens, we also explore error situations and error precisions. For our tests we simulate four systems with different performance, which represent the main capabilities of present day and future information systems:

(S1) The user is told that he is communicating with a railway employee, who does the lookup in a database and sends the information back via CRT. The system understands all utterances of the user and interprets them in an extremely cooperative manner.

(S2) System 2 behaves exactly like S1, but the user is told that he is communicating with a machine that understands unrestricted natural language and that he can question whatever he wants. S1 and S2 operate without echo.

(S3) In this system cooperativity is restricted as far as possible to perform the task in order to achieve a more machinelike behavior of the system. So the system does not take any initiatives for clarification dialogues or interprets time phrases literally.

(S4) System 4 behave like S3 but introduces additionally language restrictions. It rejects e.g. modal and vague expressions and quantifiers. Also it does not understand subsentences except simple relative clauses. S3 and S4 perform no error analysis but operate with echo.
2.2 REALISATION

For a first test, we simulated a railway information system. Forty subjects were recruited from students of the university. Some had various degrees of programming experience. They had to perform eight tasks, which took them from one to three hours. The tasks were of various difficulties and described a scenario as in the example below:

Ihre gebrechliche Großmutter möchte nach Hamburg reisen. Finden Sie eine möglichst bequeme Fahrtmöglichkeit für sie. (Your infirm grandmother wants to travel to Hamburg. Find a travelling possibility as convenient as possible for her.)

Half of the testpersons had to perform their task with keyboard input, the others had to use voice input and used the keyboard only to signal microphone on/off. So the two factors we investigated were system behaviour (with four degrees) and input modality (keyboard vs. voice). This leads to an experimental design with eight cells and five persons each.

3 RESULTS

Before starting to report the results of our study, we want to mention some difficulties appearing in our experiments. About half of the subjects in the group interacting with the operator (S1) articulated in post-experiment interviews their doubt on the human nature of their dialogue partner and vice versa about half of the subjects interacting with the human-like machine (S2) had the incorrect estimation of the interlocutor being of human nature. Nearly all of them named elements of dialog behavior which they considered proofs for being machine or human, for example identification of certain references, syntactical stereotypes and so on.

It might be a general effect of any experimental environment, that peoples distract the nature of the object in test. Therefore test designs aimed to evaluate how interlocutor models influence dialogue behavior need to be worked out carefully.

A survey of the results of the word counting and the precision procedures is given below in table 1 and 2.

The word counting procedure covered all meaningful strings of signs (inclusive of numbers and signs like äh). Time statements were counted as two (9 Uhr) or three words (9.30 Uhr).

The guidelines for counting the failures in the experimental protocols were the proper orthographical and grammatical rules in the keyboard mode and a grammar of spoken language in the voice mode.

### Table 1: values obtained for keyboard mode

<table>
<thead>
<tr>
<th>keyboard</th>
<th>type</th>
<th>token</th>
<th>TTR</th>
<th>prec</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sys1</td>
<td>152.6</td>
<td>302.2</td>
<td>0.52</td>
<td>0.899</td>
</tr>
<tr>
<td>Sys2</td>
<td>142.4</td>
<td>289.6</td>
<td>0.50</td>
<td>0.918</td>
</tr>
<tr>
<td>Sys3</td>
<td>121.2</td>
<td>343.4</td>
<td>0.36</td>
<td>0.902</td>
</tr>
<tr>
<td>Sys4</td>
<td>89.2</td>
<td>294.8</td>
<td>0.31</td>
<td>0.897</td>
</tr>
</tbody>
</table>

Statistical analysis proved, that the number of tokens is influenced by the input modes. The value of the tokens for all systems (except system 4) is in voice mode higher than the corresponding values in keyboard mode. On the other hand there does not appear significant influence of the systems on the number of tokens.

3.1 WORD COUNTING

Our hypothesis 1 (H1) that the total number of tokens is higher in communication with a human couldn’t be proved. With regard to the tokens people used, there is no difference between interacting with the human system (S1) and the human-like system (S2). This is valid for both voice and keyboard mode.

We found, that the number of types was influenced by the systems. Single comparison showed statistical difference between the following systems: (S1) : (S2, S3, S4); (S1, S2) : (S3, S4); (S2, S3) : (S4); (S1) : (S4).

The results concerning H1 are contradictory to those of similar experimental procedures done by Richards/Underwood (1984). The authors compared 32 subjects carrying out 6 tasks, among which obtaining train times from the time table service accessed over telephone. In the one case the dialogue partner was a machine, whereas in the other case it was a human being. To equalise the difference between human and synthetic voice the speech of the human partner was distorted by a vocoder. The word counting procedures done by Richards/Underwood showed a significantly higher number of words in total and a significantly higher type-token-ratio (TTR) for those subjects in the machine case.

The reason for the different results of the two studies might be found in different conditions of experimental environment, respectively in the details of the input/output–channels. Exclusively providing information on screen, as we did, may be an adequate way of providing information with computers, but it is undoubtedly an unnatural situation for human–human communication. It might therefore be sufficient for subjects to adapt their behavior more to the machine conditions than they would do in the case of being confronted with voice as an output channel.

Several other researchers compared voice mode to keyboard mode and found, that both more types and more tokens are used in voice mode. Chapapis / Parrish / Ochsman / Weeks (1977) reported data from 2-man teams solving problems cooperatively on one of four communication modes (typewriting, handwriting, voice and communication–rich mode). Zoltan/Weeks/Ford (1982) did their experiment on subjects interacting with an application program for checking account management.
Our study corresponds to these results only as far as the tokens are concerned. We found a significant difference in the total number of tokens between keyboard and voice mode. But with the statistical methods used, we couldn't find a significant difference in the number of types. The reason might be found in the experimental conditions. In the first case, concerning Chapanis et al. (1977) the experiment was done for human–human communication, whereas our experimental conditions were human–machine interaction. In the second case, concerning Zoltan et al. (1982), the experimental design is comparable to ours in the following points: It was done with a computer program, the subjects received information over screen and there were restrictions in vocabulary use. But there were two differences which might be of immense importance. The first one is the domain. The second one is task allocation, which was more detailed in the case of our experiment.

3.2 ERROR-PRECISION

The value for precision was computed by the following formula:

\[ ep = \frac{(1 - \frac{E_n}{T}) + (1 - \frac{E_p}{T})}{2} \]

where:
- \( ep \) = error-precision
- \( T \) = number of tokens
- \( E_n \) = uncorrected remaining errors
- \( E_p \) = total number of errors

Our hypothesis 2 (H2) that there is a difference in the precision of peoples’ input between the interaction with system 1 and with system 2, could not be proved either. This could also be observed for both input modes.

So, precise input doesn't seem to be a question of the subject’s interlocutor model, but of input mode. We can show that the difference in precision between the two modes is highly significant. Peoples’ failure rates increase with typing. This perception corresponds to other formal comparisons done for voice and keyboard modes (see overview in Martin (1989)).

The values we gathered for precision (summarized below in tables 1 and 2) are obviously quite homogenous.

3.3 INFLUENCE OF SYSTEM PERFORMANCE

The statistical results summarized in chapter 4.1. need further discussion. According to statistics, system differences can be observed in the number of types subjects used in interaction. In further exploratively investigating our data we found, that the different amount of types is due to significant modifications in the complexity of language and even in the structure of dialogue. I want to explain these modifications and to give some examples for it.

To begin with, looking at tab. 1, there are some striking phenomena concerning the number of tokens. Although they are not statistically relevant, they may be a hint for dialogue exploration.

The tokens-value for system 1 (spoken) relatively high. This is due to the fact that within this group single subjects’ usage of tokens extremely differed from 1136 (respectively 1055) to 224 tokens. Reasons for that are, that those subjects often talked about themes (for example that he wants to stay with his friend or go for shopping) which were not relevant to the actual task they had to ask questions about and that they frequently used channel checking utterances. Whereas the lowest value is due to the subject’s only asking one question per task. The extremely differing values in this group might allow the interpretation of being a sign of uncertainty in the interlocutor model, probably due to the computer situation. Especially the frequent use of channel checking utterances may be a hint on this. People are sensitive for this unnatural situation of communicating with a human dialogue partner. Furthermore, talking about irrelevant themes is uncustomary for human–human information-seeking dialogues on time table informations. The only dedication of this kind of information-seeking dialogue habitually is to get the required and relevant information as quickly as possible. Therefore, this behavior offends the quantitative rule for cooperation, which we observed in the dialogues collected in the FACID-Corpus (cf. Hitzenberger et al. 1986). On the other hand, the data in table 1 show, that this extreme verbosity doesn’t occur in keyboard mode. Another example for language modifications in human–machine dialogues we noticed in the subjects’ different preferences for question forms depending on the system they interacted with.

The subjects in the two groups of the non-restricted systems tended to use similar question forms to those used in human–human information-seeking dialogues. The characteristic structure of these question forms made up of a short initialisation phase (e.g. „I want ...“ + or the more complex form, as „I have a question. Could you please tell me...“), followed by the specification of place and time, and often terminate by a sentence explicitly marked as question. Although the formal structure of the question forms resemble the structure used in human–human dialogue, there a noticeable tendency of well-formed grammar in the case of usage in the context of human–machine interaction (for example certain kinds of ellipses are omitted). An example for one of the most frequently used question forms used with system 1 and system 2 is:

(a) „Ich möchte am nächsten Wochenende sehr früh am Samstag morgen von Etterzhausen nach Hindelang fahren und am darauffolgenden Tag ganz früh wieder zurück. Welche Möglichkeiten habe ich?“

(Next weekend very early on Saturday morning, I want to go from Etterzhausen to Hindelang, and I want to go back on the following day, very early in the morning. What are the possibilities I have?)

The two or even more sentences, used with this question form, contribute more to an increasing total number of types than it is the case with the question forms people use in interacting with the restricted systems. In the later form,
the questions changed to syntactically smaller constructions like simple wh-questions, for example

(b) „Wann fährt ein Zug von Hindelang nach Etterzhausen zwischen 17.30 und 22.00 Uhr?“

(When does a train from Hindelang to Etterzhausen leave between 17.30 and 22.00)

or constructions with „Is there any ...“, for example

(c) „Gibt es einen Zug, der zwischen 17.30 und 22.00 Uhr von Hindelang nach Etterzhausen fährt?“

(Is there any train between Hindelang and Etterzhausen, which leaves between 17.30 and 22.00)

Considering these conditions, one might expect that the shorter question forms like (b) or (c) might not only reduce the total number of types used, but also the total number of tokens. Although this relation seems logical, it is obviously not the case according to statistical data. A reason explaining this phenomenon is the fact that in the case of dealing with the restricted systems all ill-formed input or input contrary to the restrictions is not transmitted to the system. This has the effect, that the subjects in such cases have to repeat their question in the correct form, for which people frequently used an identical or at least nearly identical pattern, if possible. This behavior effects an increasing total number of tokens but doesn’t change the total number of types.

The dialogue protocols showed another striking phenomenon concerning the size of the subjects’ vocabulary. Those subjects consulting system 3 or 4 showed the tendency to avoid dealing with any kind of relatively extraordinary topics, such as asking for the valid conditions for using certain reduced tariffs. These topics are absolutely common to subjects in interaction with the non-restricted systems. Similarly, if the subjects indeed dealt with those topics, they didn’t use the relevant and appropriate vocabulary. To give an example for this, instead of using the term „Rückfahrt“ (return journey) the subjects consistently used their habitual question form pattern for asking for all kinds of train time informations.

That is the syntactical pattern and most of the words used with it remained, only the specifications of place and time were exchanged. One of the most frequently used forms was like (a). Considering these facts of user behavior, one can say, that the principle of questioning in the case of the restricted systems resembles the filling of slots, which resembles the usage of formal query language.

3.4 CONCLUSIONS

The statistical data discussed in chapter 4.1 and in chapter 4.2 showed, that only the number of types is influenced by the system. The number of tokens and the precision value however changes with the different input modes. For the voice mode the total number of tokens rises, whereas the number of failures falls.

In chapter 4.2 we mentioned some frequently appearing features in peoples’ verbal and dialogue behavior, (non-cooperative behavior, preference for certain question patterns) that were explored in human-machine dialogues of the non-restricted kind, which are not habitual forms in the human-machine dialogues of the FACID-Corpus. Observed these features for both interaction with system 1 and interaction with system 2 might be a hint on the nature of the users’ interaction model. The fact, that dialogue features occurring on the one hand in human-human dialogue but on the other hand no longer occur in human-machine dialogues, even not in the case of being connected with an operator, we interpret as the machine aspect of the interlocutor model is overlapping the human aspect. This is due to the medium computer.

This interpretation of explorative data is confirmed by statistical data, as demonstrated in chapter 4.1.

Furthermore, there was statistical difference found in the comparison of the non-restricted systems (S1, S2) to the restricted systems (S3, S4). This result we interprete as another change in interlocutor model. In the case of the restricted systems the quality of the interlocutor model is quite obvious, as we found immense modifications in the complexity of language (i.e. modification of question forms, avoidance of terms and even themes). With these kinds of reductions, not only the verbosity is extremely reduced, but also the coherence of the users’ dialogue acts is destroyed. Interaction with these systems is reduced to a kind of slot filling.

REFERENCES


