The Relationship Between User Errors and Perceived Usability of a Spoken Dialogue System

Antti Oulasvirta[1]*,Klaus-Peter Engelbrecht, Anthony Jameson*[2]*, and Sebastian Möller*[1]

[1]Deutsche Telekom Laboratories, TU Berlin, Germany
[2]DFKI, Germany
antti.oulasvirta@hiit.fi, klaus-peter.engelbrecht@telekom.de,
anthony.jameson@dfki.de, sebastian.moeller@telekom.de

Abstract

An experiment (N=24) was conducted with a spoken dialogue system (a smart home system), in which the users carried out several tasks with the system and rated its usability. Users' interactions were analyzed from the perspective of human error research done in human factors and cognitive ergonomics, distinguishing between goal-, concept-, task-, and command-level errors. This paper raises the question of how the interrelationship between errors and usability perceptions should be studied. Preliminary results are presented from correlational and factor-analytical approaches.

1. Introduction

An engineering approach to usability includes the consideration of how interaction with a system is perceived by the user. Although significant progress has been made in several areas affecting quality within the last two decades, there is still no consensus on the contributing components of *perceived usability*. This paper presents an empirical examination of the relationship of so-called "user errors" to the perceived usability of a spoken dialogue system (henceforth: SDS).

We here use the term "user error" broadly to refer all deviations from "optimal" interaction (see Section 3.2). Such deviations may occur when the user's expectations are not met by the system. And they sometimes pinpoint inadequacies of the system, and in such cases no fault can be given to the user. By definition, errors lead to non-progress (stagnation or regression) towards the goal of the interaction. They affect the flow of the interaction, as is reflected by quantitative measures of system performance and dialogue behavior, so-called *interaction parameters* [1][2].

An analysis of user errors may help to identify error sources and to optimize system development, as has been shown in many cases (for an example, see [3]). Reports onthe frequency and nature of errors are the basis for understanding how special and important these new interaction phenomena are in practice. Combining information on errors with information on their effects on perceived usability would help developers and designers in decision making. The present work focuses on the relationship between manually coded user errors and perceived usability as it can be measured, e.g., with the help of questionnaires (see, e.g., the SASSI questionnaire, [4], or ITU-T Rec. P.851, [5]).

2. Data Collection

The purpose of this section is to describe how the data were acquired. Since the focus of the paper is on analyzing the relationship between errors and ratings, we do not report on the experiment fully here but refer the reader to previous publications [6]. To summarize, we adopted the data from a Wizard-of-Oz experiment in which 24 participants accomplished pre-defined tasks with a multi-device smart home system called INSPIRE.

2.1. System

Some qualities of the system worth mentioning include:

- The system enables control of domestic devices (TV, video, program guide, lamps, fan, blinds, answering machine) through spoken dialogue.
- The system was installed in a laboratory decorated as a living room (at IKA, Ruhr-University Bochum).
- System output could be given from any of three loudspeaker locations.
- The user spoke to a microphone installed in the living room.
- Mixed initiation was utilized, which here means that the system tried to parse even incorrectly formulated sentences and proceed on the basis of incomplete information.

2.1.1. Experimental method

The experiment is characterized by the following attributes:

- The 24 participants (mean age 23.7 years) were recruited at the Ruhr-University Bochum, all with perfect command of the German language.
- Wizard-of-Oz simulation was used, i.e., a human replaced the system in speech recognition and typed the utterances of the user for the system.
- After an introduction to INSPIRE's capabilities, users engaged in three task scenarios, each involving 9-10 tasks with a specific device embedded in a story.
- The tasks were simple, typical tasks performed with the domestic devices; e.g., turn on/off the lamp, move the blinds up/down, switch on/off the fan, play the latest messages with the answering machine, or select a show from the evening program.
- After each scenario, users filled in a questionnaire with 37 judgments on different quality aspects.

Table 1: Mean ratings given after each interaction. Except for question 1 (overall impression: 1=bad…5=excellent), numbers indicate the agreement with the preceding statement (-2 = strongly disagree…+2 strongly agree).

No.	*Statement*	*Scale range*	*Mean*	*std*
1	Overall impression of the interaction	1...5	3.01	0.86
2.1	The system did not always do what I wanted	-2...+2	0.36	0.97
2.2	The information provided by the system was clear	-2...+2	0.89	0.76
2.3	The provided information was incomplete	-2...+2	-0.78	0.81
2.4	Home appliances can be operated efficiently using the system	-2...+2	0.06	0.85
2.5	The system is unreliable	-2...+2	-0.60	0.78
3.1	I felt well understood by the system	-2...+2	0.07	0.98
3.2	I knew at every point in time what I could say to the system	-2...+2	-0.08	1.04
3.3	I had to concentrate to acoustically understand the system	-2...+2	-1.15	0.82
3.4	The system voice sounded natural	-2...+2	0.51	0.96
4.1	The system reacted too slowly	-2...+2	0.54	0.98
4.2	The system is friendly	-2...+2	1.03	0.58
4.3	The system reacted not always as expected	-2...+2	0.58	0.90
4.4	I did not always know what the system expected from me	-2...+2	-0.14	1.10
4.5	The system frequently made errors	-2...+2	-0.63	0.85
4.6	I could easily correct errors when they occurred	-2...+2	0.53	0.90
4.7	The system reacted like a human	-2...+2	-0.93	0.91
4.8	The system behaved in a cooperative way	-2...+2	0.47	0.93
5.1	I got lost easily in the flow of the dialogue	-2...+2	-0.39	0.83
5.2	The dialogue was bumpy	-2...+2	0.39	1.00
5.3	I could control the dialogue as I wanted	-2...+2	0.23	0.84
5.4	The dialogue was too long	-2...+2	0.59	0.98
5.5	The dialogue quickly lead to the desired aim	-2...+2	-0.32	0.88
5.6	The dialogue was balanced between me and the system	-2...+2	-0.21	0.92
6.1	The interaction with the system was pleasant	-2...+2	0.30	0.94
6.2	I felt relaxed	-2...+2	0.29	0.92
6.3	I had to concentrate during the interaction	-2...+2	0.36	1.03
6.4	It was a pleasure to interact with the system	-2...+2	0.80	0.92
6.5	Overall, I am satisfied with the system	-2...+2	0.35	0.89
7.1	The system is difficult to handle	-2...+2	-0.48	0.93
7.2	System operation is easy to learn	-2...+2	1.00	0.66
7.3	Operating home appliances via speech was comfortable	-2...+2	0.50	0.90
7.4	The system was too inflexible	-2...+2	0.35	0.92
7.5	The system is not helpful for operating home appliances	-2...+2	-0.56	0.96
7.6	I prefer to operate home appliances in a different way	-2...+2	-0.06	1.10
7.7	I would use the system again in the future	-2...+2	0.56	1.01
7.8	System operation was worthwhile	-2...+2	0.36	1.01

- After the experiment, the users evaluated the whole system in terms of how well it fulfilled their expectations and what could be improved.
- The order of presentation was counterbalanced through randomization of the order of scenarios.

2.1.2. Usability questionnaire

The questionnaire filled in by the test subjects after each interaction is not a standard questionnaire, it was designed for the current experiment. Important contributions to the design process came from:

1. an evaluation experiment on a telephone-based spoken dialogue system for restaurant information (see [2] [7][9]), which lead to the standardisation of a new Recommendation on the evaluation of telephone-based spoken dialogue systems by the International Telecommunication Union (ITU-T Rec. P.851, [5])
2. the SASSI questionnaire described in Hone and Graham [4], which has mainly been designed for systems with speech input capability (not necessarily speech output).

New questions which seem to be important for a smart home system have been added, resulting in a new set of questions which should cover as broadly as possible the quality dimensions relevant for the user (see Table 1).

3. The Data Set

To summarize, the purpose of the study was to collect usability ratings and to relate them to user errors. For this purpose, a usability perception questionnaire was constructed. For the errors, a new taxonomy of user errors custom-tailored for SDS was constructed, making a distinction among 1) goal-level (i.e., misunderstanding the capabilities of the system), 2) task-level (i.e., not understanding how to reach the goal in interaction with the system), 3) representation-level (i.e., referring to the world in a way that is not understood by the system), and 4) command-level errors (i.e., vocabulary and grammar errors).

The purpose of this section is to introduce the reader to the data on ratings and errors.

3.1. Ratings

The resulting data set consists of 66 ratings for each of the 34 questions in the questionnaire (one page of the questionnaire was missing for three participants, which dropped the number from 72). Table 1 presents the ratings and their means.

3.2. Errors

The concept of an *error* is here defined in the context of goal-pursuit. Task goals are defined as transformations of objects (digital and artefactual) achievable through a hierarchy of commands given to the system (as defined by the relevant dialogue structure). For each task there is an optimal solution path, or many, “hidden” from the user’s immediate perception. The execution of the task involves commanding the system and interpreting the cues in its response so as to transform the initial state of the system to the hidden goal state. In effect, an error, then, is any deviation from the optimal solution path(s). In recording these deviations, the coding deals with overt behaviour, not with the immediate and latent reasons that cause them. Thus, the classification taps the “phenotypes” of errors (overt behaviour classifiable as an error) rather than their “genotypes” (causes and antecedents of errors).

3.2.1. Taxonomy

Table 2 presents the categorization scheme used.

The immediate effect or *consequence* of an erroneous utterance can be of three types:

1. *Stagnation.* The system takes the user to a prompt that is as close to the task goal as the previous prompt,i.e., the goal can still be reached with as many steps as before. Two special cases of this are called Repetition and Rephrasing: The system repeats the prompt (word to word or just the end of it but meaning the same thing and being pragmatically the same prompt with same action alternatives). A third special case is Help-prompt,(in which possible utterances are proposed to the user.
2. *Regression.* The system goes to a state that is farther away from the task goal than the previous state,.i.e., the user has deviated from an optimal solution path and now has to go through at least one extra state in order to achieve the goal. A special case of this is called Restart: The system returns to its initial state, losing any progress achieved in the task before the error occurred.
3. *Partial Progress.* The system goes to a state which is closer to the task goal, but not all the information in the user’s utterance is processed.

Table 2: Error categories and their definitions.

(Goal-level)

CAPABILITY $=_{df}$ The system does not posses the function or capability assumed in the otherwise valid request. Subcategories: asking the system to control 1) an object that is not in the system, 2) at a level of granularity not possible by the system, 3) in a way that is not possible due to extra-systemic restrictions.

(Task-level)

STATE $=_{df}$ Issuing a command that is valid in one state of the dialogue, but not in the current one. Subcategories: 1) progressive command valid in a future state in the optimal solution path, 2) unprogressive command valid in a previous state.

(Command-level)

VOCABULARY & GRAMMAR $=_{df}$ Issuing a command that would be valid if one word was changed to its synonym or the grammatical order of words was changed, without changing the meaning of the utterance. Subcategories: 1) word (verb, noun, adjective, adverbial) error, 2) grammatical construction error (phrasing, sentence structure).

(Concept-level)

MODELING $=_{df}$ Issuing a command that would be valid if the system represented the world in a different way. It is possible to imagine another kind of model/categorization of the world in which this utterance would not constitute an error. Subcategories: referring incorrectly to 1) time, 2) space, or 3) attribute of an object.

(Other)

OTHER $=_{df}$ All other recognizable errors. Subcategories:

1. *No input error* $=_{df}$ Failing to issue a command during the timeout interval in which the system expects it to be issued.
2. *Common ground error* $=_{df}$ Issuing a command that refers to outcomes of previous states (e.g., "Please switch on *the other* lamp")
3. *Wizard error* $=_{df}$ The wizard typed the user's command incorrectly, or there was a problem with the computer.
4. (Other)

3.2.2. Coding procedure

The unit of analysis is one exchange of information between the system and the user. For any system prompt, there is always at least one user response that lies on the optimal solution path. Because many errors can be made in any one exchange, the categories are not mutually exclusive but supportive, the goal being to characterize the nature of errors from multiple perspectives rather than from one. Several initial sessions were held for the definition of the error categories. After agreeing on the general scheme presented in Table 2, one of the authors started to code the whole data set. Five calibration sessions were held altogether in refining the categories when problematic instances appeared. After each change, the category in question was recoded in the data to ensure reliability of coding. This resulted in the final data set used in the subsequent analyses.

The data of one participant could not be analyzed because of technical problems. The final data set comprised 2343 exchanges between the system and the user.

3.2.3. Inter-rater reliability

To assess the reliability of the taxonomy, a outside coder was hired to code 300 exchanges randomly sampled from the data. The coder was trained to use the coding scheme, and several examples were provided that were not part of the to-be-coded sample. In calculating Kappa between the first and the second coder, we found that Capability, State, Vocabulary/Grammar, and Modeling errors showed substantial agreement (Cohen's Kappa over 0.60) but that there were some subtler difficulties within these classes because of small frequencies of some subtypes. For example, Time and Space Modeling errors were much less reliable than Attribute Modeling errors. The Consequence category showed much poorer agreement (Cohen's kappa falling in the range 0.186 - 0.58) than other categories.

3.2.4. Frequency of errors in the data

The coding revealed that 28% of all exchanges involved one or more errors. Of these, 40% involved a command-level error, 20% a representation-level error, 18% a task-level error, and 12% a goal-level error. (9% of the apparent errors could not be classified with our taxonomy at this point.) With 50% of the errors, dialogue flow was stagnated as a consequence of the user error; in 12% there was regression, and in 38% partial progress in the task was achieved despite the error. Some of the error types were restricted to particular devices in the system. For example, sentence construction errors were typical only in conjunction with the answering machine, spatial referencing errors only with lamps, and time referencing errors only with the program guide. In sum, user errors were very frequent at all levels of interaction, and they had serious consequences for the fluency of the dialogues.

3.2.5. Effects of different variables

A one-way ANOVA was run on the dependent variable expressing the number of errors per exchange:

- *User* had a significant effect, $F_{22,2316}$=2.57, *p*<.0001. The user with least fewest had only .15 errors per exchange and the one with the most errors .48.
- *Device* had a significant effect, $F_{6,2316}$=14.79, *p*<.0001. The blinds and the fan were associated with the fewest errors, .14 and .15 respectively per exchange, whereas the TV complex and lamps fell in the range between .32 and .44.
- *Trial number* had a significant effect, $F_{2,2320}$=8.17, *p*<.001. The third and last trial showed significantly fewer errors than the two first trials, which did not differ statistically from each other, as determined by a post-hoc test (LSD).

Moreover, there were significantly more errors, $F_{1,2321}$=5.04, *p*<.05, when the *object* of the user command was content (here: video or message) as opposed to a device, the former comprising 12% of all exchanges and averaging .40 errors per exchange.

3.2.6. Examples of error analysis

The following examples are presented here with the goal of illustrating what an analysis of errors from transcriptions looks like. The examples are quite typical cases of problematic dialogues in which several errors emerge:

> (Example 1) In this dialog, the system asks "what would you like to do with the TV show you chose?" The user's task is to signal a show's beginning. The user answers "I want to watch it". However, for the system the value "watch" means to watch immediately. Possibly he is expecting as a next system prompt something like "The show is going to be shown at 8 o'clock. Do you want me to remind you?". This does not happen, however. Following this misunderstanding, the user tries to stop the film the TV is playing by saying "Stop the film" (the correct command would be "Switch off" and "TV"). The INSPIRE system understands only "stop", which is a value valid only for the blinds, so it infers that

the blinds are to be used (this process is opaque to the user) and answers "The blinds are not moving"

(Example 2) Principally the same kind of error is found in the following attribute/verb error: The same user says "Switch on the answering machine". The system recognizes "Switch on" as the action value, which can be associated with TV, lamps etc., but not with the answering machine (which again the user does not know), so it asks the user to revise the action value "switch on" (the correct value would be "play current message"). Furthermore, when operating the lamps, the user tries 2 or 3 different synonyms for the verb "dim", none of which the system understands. The user, however, does not get any suggestion on which verb he could use. Even when a help prompt is played, it just says that the lamps, fan and TV are possible devices to switch on or off.

4. Results

The analysis of these data is still in progress. We report in this paper on correlations between individual error types and usability ratings so as to give an idea of the types of relationship that may be found when more specifically appropriate statistical techniques are applied.

4.1. Correlations

Our first analysis involved computing Pearson correlations between frequencies of errors and usability ratings. Taking the user as the unit of analysis, moderate (.5 < |r| < .8) correlations were found for all four error categories, except for goal-level errors, which were relatively rare in the data. We here report moderate correlations statistically significant with an α level of .01.

Of the task-level errors, particularly noticeable were the kinds of errors where the user repeated a command that wound have been valid after some previous prompt but was not meaningful after the current one. This error type correlated with perceptions of poor error recovery mechanisms of the system (r=.58), being stressed (r=.58), unpleasantness of interaction (r=.74), difficulties in learning (r=.74), unhelpfulness of the system (r=.58), and unwillingness to use the system again (r=.58). Of the command-level errors, noun errors were relatively frequent (14% of all errors), and they were correlated with a perception of unhelpfulness of the system (r=.66), a feeling of discomfort (r=.63), and a feeling of being lost in the dialogue structure (r=.59). Phrase errors (6.5%) correlated with a perception of a poor balance between the user and the system in the dialogue flow (r=.66). Of the representation-level errors, particularly noticeable were spatial reference errors (e.g., referring to the lamp "next to me" although the system does not know the user's location; 7% of all errors). This error type was most systematically associated with negative ratings of pleasantness, relaxedness, satisfactory performance, helpfulness, error correctability, and acceptability (all |r| > .59).

The nature of the system response when an error occurred was also correlated with usability ratings. Particularly noticeable was that a repetition of system prompts (as the response to an invalid user command) was consistently associated with negative ratings of the system's helpfulness, learnability, error recovery, reliability, and understandability (all |r| > .58). In addition, negative perception of understandability correlated with stagnation of dialogue flow (r=-.59).

Table 3: Principal component analysis with a Varimax rotation and Kaiser normalization of the ratings. Factor loadings higher than |0.60| have been marked with bold typeface.

Question		Components							
No.	Rated aspect	1	2	3	4	5	6	7	8
1	overall quality	0.49	-0.12	0.35	-0.48	0.33	-0.05	-0.17	-0.35
2.1	task failure	-0.15	-0.07	**-0.64**	0.27	-0.25	-0.05	0.03	0.30
2.2	information clear	0.24	-0.17	**0.65**	-0.28	0.16	0.04	0.32	0.04
2.3	information not complete	0.02	0.13	-0.50	0.31	-0.16	-0.58	-0.09	0.01
2.4	efficient handling	**0.63**	0.00	0.16	-0.02	0.23	0.14	0.14	-0.48
2.5	system unreliable	-0.31	0.03	0.05	**0.64**	0.05	-0.35	0.09	0.26
3.1	perceived understanding	0.50	-0.03	0.49	-0.51	0.27	0.07	-0.09	-0.13
3.2	system transparent	0.27	-0.41	0.26	0.10	0.48	0.17	-0.18	0.01
3.3	listening effort	-0.03	0.01	0.18	0.03	**-0.84**	0.16	-0.17	0.10
3.4	system voice natural	0.09	0.03	0.08	-0.22	-0.04	-0.14	**0.74**	-0.23
4.1	system slow	-0.09	0.12	-0.13	0.18	-0.09	-0.13	-0.23	**0.78**
4.2	system friendly	-0.01	-0.57	0.08	0.03	-0.09	0.31	0.18	-0.25
4.3	reaction unexpected	0.00	0.16	**-0.76**	-0.02	0.29	-0.15	0.03	0.03
4.4	syst. expectation unknown	-0.15	0.53	-0.58	-0.16	-0.07	0.25	-0.03	0.07
4.5	system made errors	-0.06	0.15	-0.27	**0.65**	-0.08	-0.01	-0.23	0.07
4.6	error recovery	0.48	-0.46	0.43	-0.02	0.28	-0.05	-0.18	0.09
4.7	human-like reaction	0.44	-0.03	0.30	0.40	0.01	0.47	0.03	-0.26
4.8	system cooperative	0.22	0.18	0.02	-0.04	-0.04	**0.77**	0.00	-0.10
5.1	loose dialogue flow	-0.37	**0.62**	-0.25	0.14	-0.16	0.02	0.18	0.09
5.2	dialogue bumpy	-0.42	0.39	-0.24	0.07	0.12	-0.13	0.18	0.44
5.3	interaction control	**0.76**	-0.08	0.19	0.07	0.08	0.06	-0.27	-0.20
5.4	dialogue too long	-0.33	0.43	-0.07	0.06	-0.07	-0.10	-0.21	0.53
5.5	dialogue fast	0.26	0.00	0.57	0.02	-0.20	0.14	0.40	-0.23
5.6	dialogue symmetric	0.23	-0.01	0.05	0.12	0.23	0.25	**0.75**	-0.05
6.1	interaction pleasant	**0.64**	-0.32	0.19	-0.24	0.12	0.28	0.28	-0.06
6.2	relaxation	0.34	**-0.60**	0.04	-0.41	0.19	0.00	0.28	-0.04
6.3	concentration required	-0.18	**0.77**	0.00	0.14	-0.06	0.15	0.04	0.07
6.4	interaction fun	0.41	-0.43	-0.06	-0.40	0.25	0.31	0.25	0.08
6.5	overall satisfied	**0.73**	-0.27	0.14	-0.19	0.26	0.10	0.18	-0.25
7.1	operation difficult	-0.48	0.32	-0.24	0.30	-0.33	-0.36	-0.04	0.07
7.2	easy learning	0.33	-0.17	0.14	-0.32	**0.62**	0.26	0.04	-0.04
7.3	operation comfortable	**0.60**	-0.20	0.10	-0.06	0.41	0.22	0.25	-0.11
7.4	system inflexible	-0.28	0.28	0.06	0.40	0.01	-0.51	-0.12	0.36
7.5	system not helpful	**-0.70**	0.26	-0.14	0.17	-0.27	-0.05	-0.20	0.10
7.6	other interface preferred	**-0.85**	0.11	0.02	0.08	0.17	-0.06	-0.10	0.05
7.7	use again	**0.82**	-0.20	0.05	-0.22	0.00	0.11	0.28	0.03
7.8	operation worthwhile	0.58	-0.22	0.16	-0.19	0.26	0.36	0.42	-0.08

4.2. Factor analysis

Another approach that we are currently exploring is the use of factor analysis to reduce the number of usability variables and hence the likelihood that high correlations can arise because of chance alone. For example, Möller et al. [6][10] conducted a factor analysis on these same data, arriving at 8 factors that can be interpreted in general terms such as *cognitive demand* and *task efficiency*. When these factors are used instead of the individual usability ratings for the computation of correlations, the highest correlations are understandably less extreme. For example, spatial reference errors (see above) show a substantial correlation only with the dimension of *acceptability*, which subsumes some of the individual variables listed above as correlating with spatial errors. The correlations with the factors appear to reflect more robust, general trends, but they may fail to capture more specific relationships between error types and individual variables.

The rest of this section explains the factor-analytic computations and their interpretation in more detail.

4.2.1. An eight-factor model for the ratings

Exploratory factor analysis was performed on the individual dialogue ratings data. Results of the principal component

Table 4: Spearman rank coefficients for the relationsihp between error types and usability rating components

	1. Acceptability	2. Cognitive Demand	3. Task Efficiency	4. System Errors	5. Ease of Use	6. Cooperativity	7. (Untitled)	8. Speed of Interaction
Error								
No input	.08	.23*	-.22*	.09	-.07	.34**	.16	.01
Capability: Magn. of control	.17	.04	-.06	-.12	-.07	-.02	.09	-.21
State	.05	-.10	-.15	-.23*	.05	.16	.02	-.11
State: Unprogressive	-.29*	-.26*	-.01	-.15	-.09	-.04	-.06	.02
Noun	-.26*	-.16	-.11	-.06	-.14	-.32**	.06	.03
Verb	-.12	-.24*	.01	-.21	-.07	-.11	-.01	-.01
Phrase	.15	-.16	.00	-.08	-.06	-.10	.18	-.02
Satzbau	-.03	.07	.18	-.09	.06	.20	.12	-.19
Modeling: Time	.01	.04	.02	.02	-.19	-.14	.05	-.23*
Modeling: Space	-.15	-.11	-.10	-.05	.13	.13	-.14	-.12
Consequence								
Stagnation	-.16	-.07	.02	-.10	-.30*	-.11	.21	.06
Repetition	-.22*	-.13	-.03	-.21	-.23*	-.05	.20	.07
Help prompt	-.03	.00	.02	.06	-.24*	-.10	.17	.01
Regression	.18	-.18	-.07	-.10	-.15	-.01	.23*	-.14
Restart	.06	-.02	-.03	-.11	-.08	.00	.19	-.20
Partial progress	.07	-.11	-.17	-.18	.05	.09	.06	-.15

* $p < .05$
** $p < .01$

analysis with a Varimax rotation and Kaiser normalization of the ratings are presented in Table 3. An 8-factor model with eigenvalues higher than 2.0 was extracted, resulting in 72.6% of the variance being covered by the cumulated factors.

The explanation of the components is as follows:

- The first dimension could be named *system acceptance*, because highest loadings are observed for the question about whether the service would be used again, helpfulness of the system, suitability of the interface, comfort, and efficiency.
- The second dimension describes *cognitive demand*, with high loadings for the required concentration and for stress.
- The third dimension is a *task-related component*, with high loadings for task success, for the clarity of the provided information, and for the transparency of the system behavior.
- Dimension 4 is related to *system errors*; it involves the frequency of system errors and the reliability of the system.
- Dimension 5 describes *the ease of use* and has two high loadings: one for listening effort and one for ease of listening.
- Dimension 6 describes the *system cooperativity*, which is the only question with a loading higher than .6 on this dimension.
- Dimension 7 shows two high loadings which cannot be interpreted in combination: the naturallness of the system voice and the symmetry of the dialogue. Both have a strong positive impact on this dimension.
- Dimension 8 characterizes *the speed of interaction* with a moderate loading for the length of the dialogue as well.

The internal consistency reliability of the question loading higher than +/-0.6 has been checked for each of the extracted eight subscales, using Cronbach's Alpha (α) as the reliability estimator. The resulting values were α =0.93 for Dimension 1, α =0.76 for Dimension 2, α =0.67 for Dimension 3, α =0.51 for Dimension 4, α=0.63 for Dimension 5, and α =0.57 for Dimension 7. Dimensions 6 and 8 contain only one question which shows a loading higher than +/-0.6. Following the argumentation outlined in [4], α values higher than 0.7 would be adequate in the early stages of scale construction; for the eight dimensions extracted here, only Dimensions 1 and 2 satisfy this criterion; however, most of the other scales only show two questions with a loading higher than +/-0.6. On all scales taken together, Cronbach's Alpha reaches α=0.95. Deletion of scales does not lead to a significant improvement of this value. However, α =0.95 can be regarded as a relatively high reliability coefficient.

Apparently, there are a large number of dimensions underlying the questionnaire. In fact, given that it was the aim of the questionnaire to get an overview of the relevant dimensions, the number of dimensions seems to indicate that this goal has been achieved with the approach presented. It can be concluded that the scales chosen for this experiment show a high reliability and that they capture a large number of quality dimensions which may be relevant for the user of a smart home system.

4.2.2. Correlations between errors and usability factors

The next step was to calculate correlations between the eight components and the error types. We averaged over the three questionnaires per user. To minimize the effect of outliers, the square root of error counts were used instead of raw numbers [11]. Spearman rank coefficient was used in addition to Pearson coefficient, because of the latter's inability to deal with the poisson distributed error counts. Infrequent error types (less than 7 occurrences) were omitted. Table 4 presents the results of this analysis and section 5.2. the most promising effects.

5. Conclusions and Future Work

We believe that systematic examinations like these may ultimately help system developers to understand which aspects of systems are worth the investment of effort in order to improve perceptions of usability.

5.1. Summary of most promising effects

To summarize, the results point toward the conclusion that an important component of usability perception emerges from user errors in interacting with a system. Different error types correlated with unique patterns of effects on usability ratings. The results also indicate that the way the system responds to an error has an important role in a user's perception of usability. The most promising effects were:

- *Low acceptability* is best predicted by the user uttering 1) an unprogressive state error (a command that would have been valid in some previous node of the dialogue but is not valid in the present one) or 2) a noun error. If the consequence of the user's error is repetition, this system response also tends to be associated with a low acceptability rating.
- *High cognitive demand* is associated with 1) no-input errors, 2) unprogressive state errors, 3) and verb errors.
- *Poor task efficiency* is associated with no-input errors.
- *Frequent system errors* are associated with state errors.
- *Ease of use* was negatively associated with undesirable consequences of errors: 1) stagnation, 2) repetition, and 3) help prompts.
- *Perceived cooperativity* was associated with 1) absence of no-input errors and 2) noun errors.

- *Speed of interaction* was, curiously enough, associated with temporal reference/modeling errors.

The questions of whether these relationships represent replicable causal connections or just incidental correlations in our data, and (in the former case) exactly what explains them, remain to be investigated in future research.

5.2. Usefulness of such results for design efforts

If we hypthetically take these relationships at face value, several implications for design emerge. For example, on the basis of our data, we would formulate the following hypotheses:

- Repeating a prompt after an interval of time may tend to reduce perceived cognitive demand and to increase perceived cooperativity (by reducing the frequency of no-input errors) but also to reduce perceived task efficiency.
- Efforts devoted to expanding the vocabulary of the system may tend to improve perceived acceptability, cooperativity, and cognitive demand (by reducing the frequency of noun and verb errors).
- Efforts to address and model the user's mental models (e.g., different conceptions of space, time, and objects) might not yield corresponding improvements in perceived usability (by reducing modeling errors).
- Making the dialogue structure "flatter", for example by reducing intermittent nodes between beginning and goal states, may tend to increase acceptability and users' appreciation of how well the system handles errors, as well as lowering cognitive demands (by reducing state errors).
- Trying to parse even erroneous commands tends to avoid stagnation, repetition, and help prompts, which may be experienced as reducing the ease of use and acceptability of the system.
- …

5.3. Pertinent problems

There are several problems that need to be tackled in this work in the future:

- An upper boundary for the correlations is determined by the reliability of the codings and ratings, which in our case could still be improved.
- Low-frequency cases might be interesting, but they do not receive enough data points for robust correlations to emerge. For example, restart is presumably more annoying as a consequence of an error than stagnation, yet this fact is not reflected in the correlations we found.
- Running many significance tests requires updating the threshold for p-values accordingly, in order to minimize the likelihood of interpreting correlations that are in fact due to chance alone.
- Factor analyses may hide important specific relationships between individual error types and usability perceptions.
- Exploratory approaches provide only correlational evidence, while ultimately we want to know the causal mechanisms producing them.

Reliable and informative statements about hypotheses such as the ones formulated in Section 5.2 will be possible only once these issues have been dealt with satisfactorily.

5.4. A multi-method approach?

In order to be better able to recognize specific relationships, we have been recently looking at relatively extreme dialogs: ones with especially low ratings with regard to at least one factor. By examining the transcript of each such dialog and the errors that occur in it, we try to understand the causes of the negative ratings in terms of the errors (or indeed any other factor that might explain them). For example, one dialog yielded especially low ratings for the factor that encompasses the variables "task success", "clarity of the provided information", and "transparency of the system's behavior". This dialog featured several sequences in which the user's request to the system was misunderstood by the system in a way that must in turn have been hard for the user to understand. This type of analysis of individual cases has the drawback of requiring some interpretation and speculation that cannot be verified; we cannot, after all, be sure of the extent to which this user's negative ratings were influenced by the errors that we see in the dialog. On the other hand, the qualitative analysis constitutes a general "reality check" and a complement to the statistical methods, which have their own interpretation difficulties. We believe that a binocular view that encompasses both types of analysis yields the most insight into the complex relationships between errors in dialog and the perceived quality of a dialog system.

6. References

[1] ITU-T Suppl. 24 to P-Series Rec., "Parameters Describing the Interaction with Spoken Dialogue Systems", *International Telecommunication Union*, Geneva, 2005.

[2] Möller, S., *Quality of Telephone-based Spoken Dialogue Systems*, Springer-Verlag, New York, 2005.

[3] Bernsen, N. O., Dybkjær, H. and Dybkjær, L., Designing *Interactive Speech Systems. From First Ideas to User Testing*. Springer Verlag, 1998.

[4] Hone, K. S. and Graham, R., "Towards a Tool for the Subjective Assessment of Speech System Interfaces (SASSI)", *Natural Language Engineering*, 6(3-4):287-303, 2000.

[5] ITU-T Rec. P.851, "Subjective Quality Evaluation of Telephone Services Based on Spoken Dialogue Systems", *International Telecommunication Union*, Geneva, 2003.

[6] Möller, S., Smeele, P., Boland, H. and Krebber, J., "Evaluating Spoken Dialogue Systems According to De-Facto Standards: A Case Study", accepted for *Computer Speech and Language*, 2006.

[7] Skowronek, J., *Entwicklung von Modellierungsansätzen zur Vorhersage der Dienstequalität bei der Interaktion mit einem natürlichsprachlichen Dialogsystem*, Diploma thesis, IKA, Ruhr-University Bochum, 2002.

[8] Möller, S. and Skowronek, J., "Einfluss von Spracherzeugung und Sprachsynthese auf die Qualität natürlichsprachlicher Dialogsysteme", in: *Fortschritte der Akustik – DAGA 2003*: Plenarvorträge und Fachbeiträge der 29. deutschen. Jahrestagung für Akustik, Aachen, Deutsche Gesellschaft für Akustik, Oldenburg, 726-727, 2003.

[9] Möller, S. and Skowronek, J., "Quantifying the Impact of System Characteristics on Perceived Quality Dimensions of a Spoken Dialogue St rvice", in: *Proceedings of the*

8th European Conference on Speech Communication and Technology (Eurospeech 2003, Switzerland) , International Speech Communication Association ISCA, CH-Geneva, Vol. 3, 1953-1956.

[10] Möller, S., “Perceptual quality dimensions of spoken dialogue systems: A review and new experimental results.” *Proceedings of Forum Acusticum 2005*, Budapest, Hungary, 2005.

[11] Hair, J.F., Anderson, R. E., Tatham, R. L., Black, W. C., *Multivariate Data Analysis (Fifth Edition)*. Prentice Hall, Upper Saddle River, NJ, 1998.