USER PREFERENCE AND CATEGORIES FOR ERROR RESPONSES IN CONVERSATIONAL USER INTERFACES

A PREPRINT

Sihan Yuan
Fraunhofer Institute for Integrated Circuits IIS
Erlangen, Am Wolfsmantel 33, 91058
cindy.sihanyuan@gmail.com

Birgit Brüggemeier
Fraunhofer Institute for Integrated Circuits IIS
Erlangen, Am Wolfsmantel 33, 91058
birgit.brueggemeier@iis.fraunhofer.de

Stefan Hillmann
Technische Universität Berlin
Berlin, Germany
stefan.hillmann@tu-berlin.de

Thilo Michael
Technische Universität Berlin
Berlin, Germany
thilo.michael@tu-berlin.de

May 24, 2020

ABSTRACT

Error messages are frequent in interactions with Conversational User Interfaces (CUI). Smart speakers respond to about every third user request with an error message. Errors can heavily affect user experience (UX) in interaction with CUI. However, there is limited research on how error responses should be formulated. In this paper, we present a system to study how people classify different categories (acknowledgement of user sentiment, acknowledgement of error and apology) of error messages, and evaluate peoples' preference of error responses with clear categories. The results indicate that if an error response has only one element (i.e. neutral acknowledgement of error, apology or sentiment), responses that acknowledge errors neutrally are preferred by participants. Moreover, we find that when interviewed, participants like error messages to include an apology, an explanation of what went wrong, and a suggestion how to fix the problem in addition to a neutral acknowledgement of an error. Our study has two main contributions: (1) our results inform about the design of error messages and (2) we present a framework for error response categorization and validation.

***K*eywords** conversational user interface · error response · user experience

1 Introduction

1.1 Background

The number of people using voice assistants is increasing [41]. While the market growth of smartphones is predicted to be at around six percent in 2021 [26], the compound annual growth rate (CAGR) of smart speakers exceeds 30% [16] and is predicted to remain at that level until the year 2023 [24, 41]. The market of chatbots has also been growing [20] and is expected to grow at a CAGR of more than 27% [23, 25]. CUI are used widely and interaction design for CUI is gaining attention (see for example [33]). Researchers and designers are working on improving usability and user experience of CUI. The design of error responses is an important factor regarding user experience, because of its heavy influence on user satisfaction [15]. Even though CUI have improved rapidly over the past few years [22], errors still inevitably occur. For instance, a study testing 4,942 queries with five smart speakers suggested that even the smartest speaker still couldn't understand over 20 percent of the queries [10], which makes well-designed error messages of CUI relevant. However, even when a computer is the source of negative emotions, it can also help alleviate them [17]. Therefore, it is sensible to design error responses in a CUI conscientiously.

CUI communicate with users via conversation [33]. Issues in conversations, like misunderstandings, are communicated differently than errors by computers. Typical technical error messages by computers include "Error", "Out of memory" and "404 not found". In human-human conversations, issues are resolved for example by apologies, that contain multiple components [21]. Conversational style can make interactions with computers more human-like [1, 2]. However, some researchers argue that CUI are in danger of trying to be too human [8]. Thus we investigate user preferences for error messages in our work.

1.2 Related Studies

Several studies have been conducted on error responses in human-computer interaction (HCI). For Graphical User Interfaces (GUI) with the traditional interaction form "windows, icons, menus, pointer" (WIMP), a study from Akgun et al. [1] shows that the use of apologetic statements makes participants feel more comfortable, respected and more sensitive about their feelings compared to non-apologetic responses. Tzeng [42] carried out a similar study and the result of the study supported the conclusion that computer apologies can help in creating more desirable experiences. Another study conducted by Park et al. [31] explored users' affective states and perceptions towards three different types of error messages in computer interfaces: apologetic, non-apologetic and neutral. The results of this study indicate that users considered the apologetic system as more usable and appealing compared to the neutral or non-apologetic system. Akgun et al. [2] conducted an experiment to explore the effect of apologetic error messages and mood states on users' moods and their self-appraisals of performance. The study indicates that the contents of apology messages can affect users' self-appraisals of their performance. When computers use the strategy of "Take on responsibility" and "Offer of repair" this is more effective than "An explanation or account". These results are consistent with those from Lewicki et al. [21], which suggests that apologies with more components are more effective than those with fewer components. Moreover, certain components such as "Offer of repair" are more important than other components, and the most important component according to Lewicki et al. is "An acknowledgement of responsibility" [21]. Although the authors study GUI, their methods and procedures for experimental design are valuable for our study. In summary, studies in related fields defined different categories for error messages [2, 21], and researchers used different measurements to evaluate UX, including users' affective response and users' self-appraisals of their performance. Results of these studies suggest users prefer error responses that give emotional support [1, 34] and include apologetic elements.

1.3 Research Questions

When studying error responses, it is common to categorize them. Previous studies assume error categories without validating if users perceive error messages to be part of these categories. However, the notion of categories of error messages is helpful, because it enables us to make comparisons across different categories to find a better solution for users. In our study, we empirically confirm the validity of categories with user judgments.

Our study focuses on users' perceptions and preferences of different categories of error responses and we investigate the following questions:

1. In a categorization by survey, how do users categorize error messages?
2. Which error response categories are preferred by users?
3. How do users want error responses to be designed?

We present our work in three stages: (1) We examine whether error responses belong to one of these three categories: acknowledgement of user sentiment, apology for error, acknowledgement of error. (2) We investigate users' preference for those error categories. (3) We conduct interviews with users and ask them what kind of error messages they prefer.

2 Method

For the first research question, we designed a categorization survey, which asked participants to categorize different error responses into three categories. For the second research question, we conducted a sorting survey to investigate participants' preferences for different categories of error responses. For the third research question, we interviewed participants about the content of error responses they prefer to hear when an error happens. Figure 1 shows a schematic overview of our study.

All participants signed a declaration of consent according to the European Union's General Data Protection Regulation (GDPR) before taking part in the experiments.

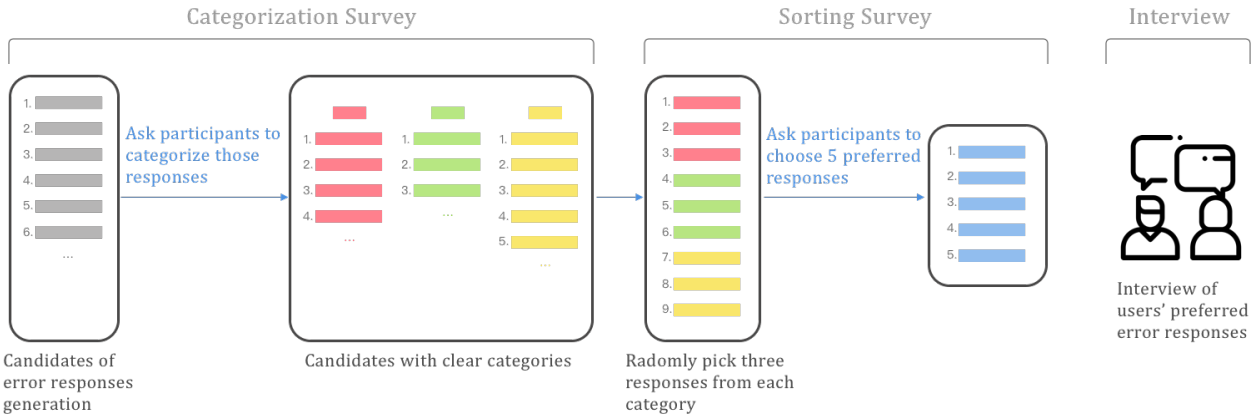

Figure 1: Visualization of our experimental design. We first created candidates for error responses. Then we presented the candidates to participants and asked them to assign them to one of three categories. We selected only candidates that were assigned by a majority of participants to the same category (for more information on our criteria for including response candidates, please refer to Section 2.1). We then conducted a second survey (see Section 2.2) in which we asked participants to rank error response candidates according to their preference. Finally, we conducted interviews with users of speech assistance systems, asking them how they want error responses to be designed. (Icons made by Freepik from `www.flatiom.com`)

2.1 Categorization Survey

We empirically confirmed the validity of assigning error responses to one of the three categories (acknowledgement of user sentiment, acknowledgement of error and apology) by asking participants to select the best fitting category for each error response candidate.

2.1.1 Response Candidate Generation

Other HCI researchers have applied apologetic strategies from HHI to HCI design [2]. We do this too and apply components of effective apology in HHI [21] to the generation of candidate error responses. According to Park et al. [31], systems that use apologetic error responses are perceived as more usable than systems that do not apologize and are neutral. Hence we decided to include the categories of apology and neutral acknowledgement in our study. Moreover, neutral error responses are commonly used in GUIs when they show for example "system error", or "404 not found". As neutral error responses are common in HCI, we investigate them as the second category of responses.

The third category of error responses was acknowledging user sentiment. Automatic emotion recognition is investigated by other HCI researchers [36] and applied in industry [5]. One reason may be that computers that respond to users' negative emotions may alleviate frustration [31, 34]. Thus we included the category "Acknowledging user sentiment" in our study.

We cover three categories of error responses in our study:

1. Acknowledging user sentiment in response to errors
2. Apologizing for errors
3. Acknowledging errors

Acknowledging user sentiment (abbreviated as: "Sentiment") means that the system response indicates empathy to users' emotions or feelings. Candidate responses include "Are you alright?" and "I know how you feel". The category "Apologizing for errors" (shortened to :"Apology") includes candidate responses like "I am sorry" or "Sorry". The category "Acknowledging errors" (shortened to: "Neutral") represents a neutral attitude like in "system error". For each category, we generated candidates that belong to the category based on our perception: for the "Acknowledging user sentiment" category, we generated 15 candidates, for the category of "Apologizing for errors" we came up with 18 responses and for the category "Acknowledging errors" we created 11 candidates. In total we thus presented 44 error response candidates to participants. The candidate list was proofread and edited by all authors, still generation of response candidate is subjective.

2.1.2 Participants

The study required participants' fluency in English since the survey was in English. Besides good understanding of English, there was no other requirement for the participants. We did not ask participants to provide demographic information for this survey, as we attempt to minimize the collection of sensitive data as part of our institute's privacy policy. Essentially, we weighed the collection of sensitive information like gender and age with expected effects of these demographic variables on our dependent variable: perception of categories. We expected people's gender and age to not affect their perception of error response categories; therefore, we decided not to collect demographic information for this survey. We designed and distributed the internal survey within our institute and received 61 responses in total.

2.1.3 Survey Design

We used a hermeneutic approach [32] for validating assignments of categories to error responses. This means we first generated candidates for each of the three categories, then we presented those candidate error responses to participants. We designed the survey with Limesurvey (`https://www.limesurvey.org/`) and participants were able to access it online. At the beginning of the survey, the task was explained in writing to the participants, along with the definition of each category. We designed the survey using multiple-choice items, and we asked participants to choose one option for each question. Each category was displayed as an option for participants to choose from. We asked them which of the three categories they chose as most fitting for every candidate. Then we analyzed the categorization responses and selected only those responses that belong clearly to one category as rated by the users. This process redefines which error responses we consider to be members of a category.

2.1.4 Data Analysis

We filtered error response candidates based on the agreement of participants about assigning responses to a certain category. To show that participants agree on categorization, we required evidence that a majority of them categorize a response in a similar way [37]. Therefore, we use Krippendorff's alpha (α), a metric that measures inter-rater agreement [4, 37]. In our study, the term "rater" refers to participants who took part in the categorization survey. Alpha (α) is given by:

$$\alpha = 1 - \frac{D_o}{D_e} \tag{1}$$

Where D_o is the disagreement observed and D_e is the expected disagreement. Expected agreement refers to how much agreement one expects by chance. For example, if two raters were to assign items to two categories, we expect them to agree in 50% of cases by chance. The smaller the observed disagreement (D_o), the higher the agreement among raters. Researchers generally consider $\alpha > 0.8$ to signify high agreement, and $0.8 \geq \alpha > 0.667$ to allow tentative conclusions of inter-rater agreement [19].

2.2 Sorting Survey

The categorization study, described in Section 2.1, resulted in error responses that showed at least a tentative agreement between raters. We used error responses selected in the categorization survey as basis for our sorting study. The results of the sorting study indicate participants' preference towards error response categories. Additionally, the sorting study allows the ranking of individual error responses.

2.2.1 Survey Setup

The sorting survey was designed to consist of two steps. First, participants were asked to choose their preferred five out of nine error responses. In the second step, participants were asked to rank these five responses. Miller's theory [28] says that the number of items people can hold in their short-term memory is seven plus or minus two. We assume that effective ranking requires participants to remember items, which is why we asked participants to rank no more than five items. We decided to present nine items in the first step, so that we can present three error responses from each category for every participant. The survey was internal and conducted within our institute, and we collected 91 responses. For this survey, we also weighed privacy considerations as well as scientific value and we decided to collect demographic information as we believe that preferences may be affected by age or gender. Of the 91 participants, 58 of them are male and 32 of them are female. 13 of the participants are between 18 and 24 years old, the majority of participants, 60 in total, is between 25 and 39 years old and 17 participants belong to the age group ranging from 40 to 59 years. One participant did not provide their age and gender.

2.2.2 Method of Analysis

Each participant ranked responses from 1st to 5th place according to their preference. We converted rankings to points in our analysis: 1st place is converted to 5 points, 2nd place to 4 points, 3rd place to 3 points, 4th place to 2 points and 5th place is assigned 1 point. When a response is not chosen to be in the top five (out of nine) the response earns no points. Equation 2 shows the formula for calculating the sum of points P_{sum} for a response across participants.

$$P_{sum} = \sum_{1}^{5} \left(\mathrm{n}^i * (6 - i) \right) \tag{2}$$

Here n^i refers to the number of times a response is ranked on place i, where $5 \geq i \geq 1$. For instance, if a response was chosen by one participant only and then ranked first, its sum of points would be: 5 points $\times$ 1 time = 5.

Also, we define D_t as the number of times a response was displayed, i.e. presented to a participant. We randomized error responses using the "random order" function in Limesurvey which converges to equal presentation probabilities for all items over all trials. We randomly selected three responses from each category, thus nine responses in total were displayed to each participant. Hence the number of times a response was displayed to participants may vary. Moreover, we define C_t as the number of times a response was chosen by participants as one of the top five out of nine displayed responses. By considering D_t and C_t we avoid misinterpreting total points P, that may be interpreted differently relative to D_t and C_t. In our analyses we consider three metrics: (1) Percentage of times a response was chosen $C_{\%} = \frac{C_t}{D_t}$, (2) Preference relative to times a response was displayed $D_p = \frac{P}{D_t}$ and (3) Preference relative to times a response was chosen $C_p = \frac{P}{C_t}$.

We used Mann-Whitney U to test for significant differences in preference between error categories. Mann-Whitney U test is a non-parametric test that does not require normal distribution of data [29]. Because there were two comparisons of each category for each metric, we used the Bonferroni-Holm procedure [14] to correct for multiple comparisons.

2.3 Interviews

We interviewed 30 people asking them "If there is an error when interacting with a voice user interface, what kind of error message would you like to hear?" The recruitment was both internal and external, twenty males and ten females were interviewed. 28 of the interviewees are between 18 and 39 years old and two participants are between 40 and 59 years old. With the interviews, we aim to generate insights from a qualitative perspective. We recorded interviews using a Zoom H1 Handy Recorder and took notes during the interview. Recordings were transcribed by both the experimenter and Google Cloud Speech-to-Text to avoid potential transcribing errors. We analyzed interview data using content analysis, which involves systematically coding data to discover patterns [11]. One of the author generated a first draft of the content analysis book. The categories and rules in this draft were then reviewed and amended by another co-author. We did not conduct the content analysis multiple times with independent raters.

The coding of qualitative data involves two stages. Firstly, the experimenter separates qualitative material into units, secondly, category-sets are established that represent patterns in the data and can be used to establish a theory afterward [12, 38]. Our procedure involved three stages: (1) We identified answers that contain some expressions, which were brought up repeatedly during the interview. (2) We categorized answers according to those expressions. (3) We named the categories. The data collected from each participant was labeled and assigned to several categories according to our best judgment.

3 Results

3.1 Categorization Survey

Exemplary, Table 1 shows a part of the results of the categorization survey. All error response candidates as well as the results of the categorization survey for each candidate are provided online [44]. In the data, agreement for error responses varies from 1 to -0.0081, where 1 is perfect agreement across participants and values around 0 indicate that raters agreed at chance level.

Among the 44 error response candidates, 16 of them had an α greater than 0.8, which is considered to indicate high agreement [18]. Only one response from the category "apologizing for errors" was above the threshold of $\alpha = 0.8$, i.e. "I'm sorry". When we set the cutoff α to 0.68, which indicates tentative agreement, this resulted in at least four candidates for each error response category. Following this criterion ($\alpha = 0.68$), eight responses from category "Sentiment", nine responses from the "Neutral" category and four responses from the category "Apology" are qualified as sufficiently clear.

Error Messages	α	Category
Currently I am not able to process your request.	1	Neutral
You seem to be upset.	0.951	Sentiment
System error.	0.951	Neutral
An error has occurred.	0.951	Neutral
Are you alright?	0.903	Sentiment

Table 1: Top five error responses with the highest inter-rater agreement α.

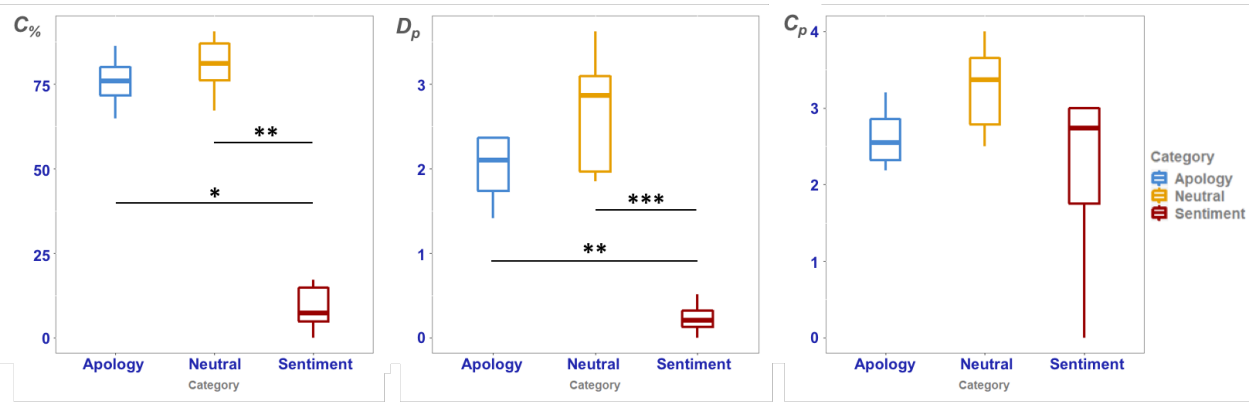

Figure 2: Boxplots for the result of the sorting survey under different criteria with value of median, upper and lower quartiles. Different colors represent different categories of predefined error messages, and three panels represent the result from different metrics: (a) Preference of times a response was chosen ($C_{\%}$). (b) Preference relative to times a response was displayed (D_p). (c) Percentage of times a response was chosen (C_p). The legend on the right refers to all three panels. The demonstration of significance of difference across three categories are as followed: 0.005<p<0.025: *, 0.0005<p<0.005: **, p<0.0005: ***.

3.2 Sorting Survey

In the sorting survey, participants were asked to choose five out of nine error responses they liked most, and then rank these five responses.

In Figure 2 we present results of participants' preferences for the three error response categories (Neutral, Sentiment, Apology) using the three metrics described in the related methods section (2.2.2): (1) Percentage of times a response was chosen $C_{\%}$, (2) preference relative to times a response was displayed D_p and (3) preference relative to times a response was chosen C_p. The higher the value of a metric, the higher participants' preference of an error response category. Figure 2 shows that after Bonferroni correction, two of the three metrics (preference relative to times a response was displayed D_p and percentage of times a response was chosen $C_{\%}$) indicate significant differences between sentiment and non-sentiment categories of error responses. The metric preference relative to times a response was chosen C_p showed no significant difference between categories.

3.3 Interviews

As shown in Table 2, our content analysis of interviews identified three categories of responses. Our interviewees wanted error messages to (1) explain the reason for the problem, (2) apologize for the error and (3) provide suggestions or guidance. The number in brackets in Table 2 indicates the number of times those elements are mentioned by participants. Participants were able to mention more than one element, which is why the number of mentions exceeds the number of participants ($n = 30$). The element "Explain the reason for the problem" was requested most frequently, being mentioned 16 times. The second most frequently requested element of error responses was suggestions and guidance by the system, which was mentioned eleven times. Descriptions included "I don't know what to do" or "Perhaps tell me what to do". Moreover, participants expressed that they want the system to be more polite and apologize. This was mentioned five times, which suggests that participants like error messages to include an apology.

Initial keywords	Modified category	Number of mentions
Specify the error Tell me the problem More detail	Explain the reason for the problem	16
Kind response Be polite Sorry, ...	Apologize for the error	5
Could you repeat that again? Please try again Indicate the state	Provide suggestion or guidance	11

Table 2: Content analysis of the question "If there is an error, what kind of error messages do you like to hear?". The column *Initial keywords* displays the keywords we used for our content analysis. The column *Number of mentions* indicates the number of times those elements were mentioned by participants.

4 Discussion

4.1 In a categorization by survey, how do users categorize error messages?

Results of the categorization survey show that participants agree most on categorizing error responses from either the neutral or the sentiment category. In contrast, most error response candidates from the apology group show relatively low agreement (below an α of 0.8). A reason for this may be that an expression of apology can mean more in human-human interaction than just an apology. For instance, "I'm sorry" can be perceived to express regret [21], but can also be used to show empathy [39]. Moreover, some participants may have perceived simple apologetic responses as default answers of system failure. This might be due to websites showing apologetic messages when something goes wrong and users learning to expect this response as a neutral acknowledgement of an error. Another interpretation of an apology can be expressing empathy. In fact, 7 out of 61 participants considered "Sorry" as either neutral or acknowledging user sentiment.

The response "I promise I will figure it out" had the lowest agreement score of all responses, with α = -0.0081. 24 participants considered it as neutral, 18 participants thought it indicated sentiment and 19 participants categorized it as apology. This might be explained by "promise" being an expression which is usually used in human conversations [3, 40] and can be interpreted in various ways [35]. The sentence can be understood as a way to express regret, a commitment [35] to take on responsibility or an indirect indication that there was an error.

4.2 Which error response categories are preferred by users?

To the best of our knowledge, there are no published methods for analyzing a two-stage sorting experiment as conducted by us. Thus we decided for three measures ($C_{\%}$, C_p and D_p) to measure preference for error response categories in our set-up. We designed the metrics in such a way as to avoid bias caused by single-criterion measurement. The criteria percentage of chosen ($C_{\%}$) and preference of displayed (D_p) show significant differences between the sentiment group and the other two groups. However, the criterion preference of chosen (C_p) shows no significant difference between categories. One possible explanation is that those three criteria measure different aspects of participants' preferences. Both $C_{\%}$ and D_p show a similar pattern of results: non-sentimental responses are preferred over-sentimental responses. Hence both $C_{\%}$ and D_p seem to measure a similar construct. In contrast, the criterion preference of chosen (C_p) shows different results, finding no difference between categories. This may be explained by how C_p is computed. Similarly to D_p, C_p takes into account ranking of the short-listed top five response. Contrary to D_p however ranking scores are analyzed relative to the number of times responses were short-listed, rather than displayed. This means, that once a participant short-listed a response there was no significant difference between their preferences. This suggests that error responses belonging to the sentiment category were rarely short-listed which is why $C_{\%}$ and D_p scores are lower for this category than for the other two categories. However, when participants short-listed error responses that included sentiment, they tended to like them just as much as neutral or apologetic responses. $C_{\%}$ and D_p seem to measure how many people like a category and C_p measures how much people like a category. That means this criterion (C_p) may be sensitive to identify small groups of users with special preferences.

The results on preferences for error response categories show the tendency that if an error response has only one element (neutral acknowledgement of error, apology or sentiment), responses that acknowledge neutrally that an error happened are preferred by participants. Therefore, we suggest to use error responses with neutral content for default design.

4.3 How do users want error responses to be designed?

In the sorting study, we explored participants' preferences of error responses containing one element. However, error messages may can contain more than one element and the multitude of elements may improve the user's experience of error messages. We interviewed participants and asked them what elements they expect in an error messages, and three response elements were brought up repeatedly: explaining the reason for the error, providing guidance and apologizing for the error. Our sorting survey indicates no significant difference in preferences for neutral and apologetic messages, which validates that apologetic elements are perceived as important for error responses by users. Unlike the sorting survey, the interview question encouraged participants to come up with elements for error messages themselves, rather than choosing from predefined ones. The interviews suggest that users prefer error responses of CUIs with apologetic elements, explanations of reasons for the error and suggestions or guidance by the system. Note that "explanation of reasons" contains "neutral acknowledgement of errors", and adds an explanation to it.

Our results suggest that participants express different preferences when asked in a survey and an interview. In the survey, participants reported preferences for neutral error messages, with one element only, while in the interviews error messages were preferred that included several elements, including apologies and guidance. One explanation is that in the survey, participants were limited in their responses and could select only between three groups of single-element responses. In contrast, in the interview we asked about preferences in an open question. Participants were able to provide feedback more freely, naming error response elements that were not covered in the survey (e.g. guidance), and expressing preferences for error messages with multiple elements.

4.4 Limitations

4.4.1 Candidate generation

Researchers suggest that an inter-rater agreement of $\alpha > 0.8$ is acceptable [30, 37]. Applying this threshold to the category "Apology" rendered only one candidate out of 15 acceptable. As a single error response is not sufficient to represent a complete category, we decided to include a total of four error responses from the category "Apology". For this we included error responses with $\alpha = 0.8$ and we set the cutoff to $\alpha = 0.68$. Hence the error responses that we included with an $0.8 \geq \alpha > 0.68$ may not be sufficiently representative for the category. This problem can be solved in future work by generating more candidates for each category. Additionally, our results suggest that users disagree in their perception of apologetic elements and it will be interesting to further study the perception of "Apology" in error messages. Our interviews show that users want to get apologies for errors, and our categorization study demonstrates that even a seemingly clear apology in the form of "Sorry" is perceived by many people as a neutral acknowledgement or sentiment rather than as an apology. This raises the question of how a system should apologize to users in a way that they perceive it as an apology.

4.4.2 Biased sample of participants

Our two surveys were distributed within a research institute, and employees there are mostly engineers. Engineers may show the tendency to be more rational and realistic than other professions [7, 13], which may cause biased results. For example, engineers tend to make decisions based on logic, fact, and data versus feelings [43], therefore neutral responses which indicate the error may be preferred more by engineers than non-engineers. Furthermore, rationality of engineers may cause bias when rating emotional contents. To eliminate this possible bias, a wider range of people should be surveyed.

4.4.3 Lack of demographic information

We followed the principle of data reduction and data economy for privacy reasons. Therefore we decided to not collect certain demographic information in the categorization study, because we we have no indication that people's perception of categories might be influenced by their gender and age. However, to the best of our knowledge, there is no research showing that people's perception of error response categories is not affected by age and gender and thus this assumption is a limitation of our work.

4.4.4 Content analysis

For the content analysis two authors compiled the code book. The content analysis, however, was conducted by one rater only. This may limit the validity of our analysis [9].

4.5 Future Work

Our current work is based on surveys and interviews and we suggest validation of our results, for example by replicating this work with a more diverse sample of participants, specificially a less engineer-biased sample of participants. Moreover, the validated error responses from our study can be used to investigate error responses in real interactions with CUI in future research, for instance using a Wizard-of-Oz approach [6]. Our work does not replace analyses of real interactions, but can act as an entry into more in-depth analysis of error responses in CUI.

In our study, we explored the content of error responses, but we did not investigate how error response should be communicated. For example, error messages can be communicated using visual and auditory stimuli. We believe that *how* an error is communicated is at least as important as *what* is communicated [27]. Thus, future research should investigate how errors should be communicated by CUIs.

Our study indicates that a small group of users exists that likes error responses with sentimental elements. We did not probe the validity of this finding or explore why people have this preference. Thus investigating preferences of sentimental elements in error responses requires further research.

Acknowledgments

This work has been supported by the SPEAKER project (01MK20011A), funded by the German Federal Ministry for Economic Affairs and Energy.

References

[1] Mahir Akgun, Kursat Cagiltay, and Jeng-Yi Tzeng. Computer apology: the effect of the apologetic feedback on users in computerized environment. In *Fifth IEEE International Conference on Advanced Learning Technologies (ICALT'05)*, volume 2005, pages 254–256. IEEE, 2005.

[2] Mahir Akgun, Kursat Cagiltay, and Deniz Zeyrek. The effect of apologetic error messages and mood states on computer users' self-appraisal of performance. *Journal of Pragmatics*, 42(9):2430–2448, sep 2010.

[3] Pall S. Ardal. "And That's a Promise". *The Philosophical Quarterly*, 18(72):225–237, jul 1968.

[4] Ron Artstein and Massimo Poesio. Survey Article: Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):1–42, 2009.

[5] Audeering. Emotion detection in real time, 2020.

[6] Birgit Brüggemeier and Philip Lalone. WoS – Open source wizard of oz for speech systems. *CEUR Workshop Proceedings*, 2327, 2019.

[7] Luiz Fernando Capretz. Personality types in software engineering. *International Journal of Human Computer Studies*, 58(2):207–214, feb 2003.

[8] Leigh Clark and Benjamin Cowan. Voice assistant technology is in danger of trying to be too human, 2019.

[9] Philip R Doyle, Justin Edwards, Odile Dumbleton, Leigh Clark, and Benjiamin R. Cowan. Mapping perceptions of humanness in speech-based intelligent personal assistant interaction. *arXiv preprint arXiv:1907.11585, 2019.*, 2019.

[10] Eric Enge. RATING THE SMARTS OF THE DIGITAL PERSONAL ASSISTANTS IN 2018, 2018.

[11] Debra A. Friedman. How to Collect and Analyze Qualitative Data. In *Research Methods in Second Language Acquisition*, pages 180–200. John Wiley & Sons, Ltd, Chichester, UK, feb 2012.

[12] Harold Guetzkow. Unitizing and categorizing problems in coding qualitative data. *Journal of Clinical Psychology*, 6(1):47–58, jan 1950.

[13] Ross Harrison, Don T. Tomblen, and Theodore A. Jackson. Profile of the Mechanical Engineer III. Personality. *Personnel Psychology*, 8(4):469–490, 1955.

[14] Sture Holm. Board of the Foundation of the Scandinavian Journal of Statistics A Simple Sequentially Rejective Multiple Test Procedure A Simple Sequentially Rejective Multiple Test Procedure. *Source: Scandinavian Journal of Statistics Scand J Statist*, 6(6):65–70, 1979.

[15] Jiepu Jiang, Wei Jeng, and Daqing He. How do users respond to voice input errors? In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval - SIGIR '13*, page 143, New York, New York, USA, 2013. ACM Press.

[16] Bret Kinsella. Smart Speaker Sales to Rise 35% Globally in 2019 to 92 Million Units, 15 Million in China, Growth Slows - Voicebot.ai, 2019.

[17] Jonathan Klein, Youngme Moon, and Rosalind W. Picard. This computer responds to user frustration. In *CHI '99 extended abstracts on Human factors in computing systems - CHI '99*, volume 14, page 242, New York, New York, USA, 1999. ACM Press.

[18] Klaus Krippendorff. *Content Analysis: An Introduction to Its Methodology (second edition)*. Sage publications, 2004.

[19] Klaus Krippendorff. Reliability in Content Analysis. *Human Communication Research*, 30(3):411–433, 2004.

[20] Karolina Kuligowska. Commercial Chatbot: Performance Evaluation, Usability Metrics and Quality Standards of Embodied Conversational Agents. *Professionals Center for Business Research*, 2(02):1–16, 2015.

[21] Roy J. Lewicki, Beth Polin, and Robert B. Lount. An Exploration of the Structure of Effective Apologies. *Negotiation and Conflict Management Research*, 9(2):177–196, 2016.

[22] Loupventures. Annual Digital Assistant IQ Test – Siri, Google Assistant, Alexa, Cortana., 2018.

[23] Market Research Engine. Chatbot Market By Platform Analysis (Web-based, Mobile, Stand-alone); By Enterprise size Analysis (Small Enterprises, Medium Enterprises, Large Enterprises) and By Regional Analysis – Global Forecast by 2018 - 2024 | Marketresearch, 2020.

[24] Marketsandmarkets. Smart Manufacturing Market Size, Growth, Trend and Forecast to 2023, 2018.

[25] MarketsandMarkets. Chatbot Market Worth $9.4 Billion by 2024 - Exclusive Report by MarketsandMarkets, 2019.

[26] Marketsandmarkets. Smartphone Market | Size, Share | Growth, Trends | Industry Analysis | Forecast | Technavio, 2020.

[27] Marshall McLuhan and Quentin Fiore. The medium is the message. *New York*, 123:126–128, 1967.

[28] George A. Miller. The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review*, 63(2):81–97, 1956.

[29] Nadim Nachar. The Mann-Whitney U: A Test for Assessing Whether Two Independent Samples Come from the Same Distribution. *Tutorials in Quantitative Methods for Psychology*, 4(1):13–20, 2016.

[30] KA Neuendorf. *The content analysis guidebook*. SAGE Publications, Inc., 2016.

[31] S. Joon Park, Craig M. MacDonald, and Michael Khoo. Do you care if a computer says sorry? In *Proceedings of the Designing Interactive Systems Conference on - DIS '12*, page 731, New York, New York, USA, 2012. ACM Press.

[32] Margo Paterson and Joy Higgs. Using Hermeneutics as a Qualitative Research Approach in Professional Practice. *The Qualitative Report*, 10(2):339–357, 2005.

[33] C Pearl. *Designing voice user interfaces: principles of conversational experiences*. O'Reilly Media, Inc., 2016.

[34] Rosalind W. Picard and Jonathan Klein. Computers that recognise and respond to user emotion: Theoretical and practical implications. *Interacting with Computers*, 14(2):141–169, 2002.

[35] Denise M. Rousseau. Schema, promise and mutuality: The building blocks of the psychological contract. *Journal of Occupational and Organizational Psychology*, 74(4):511–541, nov 2001.

[36] Nicu Sebe, Ira Cohen, Theo Gevers, and Thomas S. Huang. Emotion recognition based on joint visual and audio cues. In *Proceedings - International Conference on Pattern Recognition*, volume 1, pages 1136–1139. IEEE, 2006.

[37] Mack Shelley and Klaus Krippendorff. Content Analysis: An Introduction to its Methodology. *Journal of the American Statistical Association*, 79(385):240, mar 1984.

[38] Joanna Smith and Jill Firth. Qualitative data analysis: the framework approach. *Nurse Researcher*, 18(2):52–62, jan 2011.

[39] Yulia A. Strekalova, Janice L. Krieger, A.J. Kleinheksel, and Aaron Kotranza. Empathic Communication in Virtual Education for Nursing Students. *Nurse Educator*, 42(1):18–22, jan 2017.

[40] T Swan and OJ Westvik. *Modality in Germanic languages: historical and comparative perspectives*, volume 99. Walter de Gruyter, 1996.

[41] Tractica. Voice and Speech Recognition | Tractica, 2018.

[42] Jeng-Yi Tzeng. Toward a more civilized design: studying the effects of computers that apologize. *International Journal of Human-Computer Studies*, 61(3):319–345, sep 2004.

[43] Jeanine M Williamson, John W Lounsbury, and Lee D. Han. Key personality traits of engineers for innovation and technology development. *Journal of Engineering and Technology Management*, 30(2):157–168, 2013.

[44] Sihan Yuan, Birgit Brüggemeier, Stefan Hillmann, and Thilo Michael. Complete Result of Categorization Survey, 2020. Commit a5ef2b4.