Mixed-Modality Interaction in Conversational Recommender Systems

Yuan Ma, Timm Kleemann and Jürgen Ziegler

University of Duisburg-Essen, Duisburg, Germany

Abstract

Recent advances in natural language processing have made modern chatbots and Conversational Recommender Systems (CRS) increasingly intelligent, enabling them to handle more complex user inputs. Still, the interaction with a CRS is often tedious and error-prone. Especially when using written text as the form of conversation, the interaction is often less efficient in comparison to conventional GUI-style interaction. To keep the flexibility and mixed-initiative style of language-based conversation while leveraging the efficiency and simplicity of interacting through graphical widgets, we investigate the design space of integrating GUI elements into text-based conversations. While simple response buttons have already been used in chatbots, the full range of such mixed-modality interactions has not yet been investigated in existing research. We propose two design dimensions along which integrations can be defined and analyze their applicability for preference elicitation and for critiquing the CRS's responses at different levels. We report a user study in which we investigated user preferences and perceived usability of different techniques based on video prototypes.

Keywords

conversational recommender systems, user interface, preference elicitation, critique-based recommendations

1. Introduction

In recent years, conversational styles of interaction have increasingly been applied in the field of recommender systems. Conversational Recommender Systems (CRS) aim to provide a more human-like and more comprehensible form of eliciting users' preferences and recommending suitable items [1]. While many e-commerce sites present recommendations in a static fashion allowing little or no user interaction, the need for more flexible and personalized ways of providing recommendations is increasingly recognized. For this purpose, CRS are utilized, whereby a virtual agent interacts with the user [2]. A variety of techniques has been explored for providing conversational interaction with a recommender, some systems, for example, follow a strict rule-based process requiring the user to answer questions with predefined answers [3], while other approaches are designed to mimic a natural conversation in which users can freely formulate their questions and answers [4].

Rather than providing one-shot recommendations with only limited user intervention, CRS enable users to respond to the recommendations they receive, to criticize them, or to provide

IntRS'21: Joint Workshop on Interfaces and Human Decision Making for Recommender Systems, September 25, 2021, Virtual Event

yuan.ma@uni-due.de (Y. Ma); timm.kleemann@uni-due.de (T. Kleemann); juergen.ziegler@uni-due.de (J. Ziegler)

CEUR Workshop Proceedings (CEUR-WS.org)

more precise indications of their own preferences in interactive conversations with the virtual agent [4]. The mode of interaction in current CRS is predominantly text-based where textual input by the user is followed by textual responses. In some cases, however, CRS also offer the user a set of potential answers in a GUI style, presenting the options as buttons. Augmenting the text-only interaction by GUI elements serves two purposes: (1) it provides users with a more efficient input technique and (2) it reduces the number of misrecognitions which may occur when just relying on typed (or spoken) input. While combining textual and GUI interaction in a CRS can increase its effectiveness and usability, the design space of possible multimodal CRS interactions has not yet been sufficiently explored yet and there is a number of options that have not been investigated or even considered in CRS research.

In this paper, we aim at providing a first, more complete investigation of the different ways how textual interaction can be combined with GUI-like interactions. Keeping the flexibility of free text interaction, we introduce and investigate a variety of additional interactions by which the textual interaction may be augmented. The options investigated include directly changing (critiquing) features of an item shown as recommendation in the dialog as well as interaction with the textual responses given by the system.

We developed video prototypes for these interactions and investigated them in an online user study. The results provide initial insights into interaction methods that users may prefer.

2. Related Work

CRS exhibit a number of potential advantages over conventional GUI-based recommenders. They can provide a more natural and less obtrusive way of obtaining information about the user's preferences which is essential for generating personalized recommendations [5]. In contrast to upfront elicitation steps that are detached from the actual recommending, such as rating a number of sample items e.g. MovieLens,[1] or completing initial interviews [6] or personality questionnaires [7, 8], the elicitation of needs and preferences can be smoothly integrated into the dialog flow, thus also mitigating the cold-start problem. Provided a sufficient level of language understanding on the part of the system, the expression of user intentions [9], preferences or even dislikes is more flexible in comparison to system-initiated GUI interactions and closed-form questions. This flexibility may come, however, at the cost of lower efficiency, especially when users need to type their questions and responses instead of just clicking one of several pre-defined options. Therefore, CRS should aim to achieve an acceptable flexibility-efficiency trade-off [10].

A further, and so far less considered aspect of CRS is their capability to provide means for critiquing the system's recommendations, thus increasing user control over the recommendations. According to Chen and Pu [11] three critiquing approaches can be distinguished: natural language dialog-based critiquing (NLC), system-suggested critiquing (SC) and user-initiated critiquing (UC). NLC as a specific form of conversation, either text-based or voice-based, is well compatible with the general CRS approach. NLC can be performed in a human-like style, simulating, for instance, the conversation with a salesperson (e.g. ExpertClerk [12]). SC on the other hand provides system-initiated critiquing options and asks users for one or more

[1] http://www.movielens.org

responses, such as, for example, in multi-attribute utility theory (MAUT)-based Compound Critiques [13]. The benefits of SC lie mainly in their ability of guiding users concerning relevant and acceptable feedback criteria, so that the system can better understand users and enhance its recommendation effectiveness. UC offers users a more user-initiated form of critiquing, such as in Example Critiquing [14]. The main advantage of UC is that it allows for a higher level of user control. Also, hybrid critiquing techniques have been proposed [11] and compared [15]. By means of dialogs in CRS, both UC and SC may be combined and used flexibly. SC and UC are often realized using graphical user interface elements, such as buttons, sliders and checkboxes to either respond to system questions (SC) or to change properties of a recommended item (UC).

There exist a large and increasing variety of techniques for realizing CRS [2]. Recent approaches for improving CRS performance include, among others, knowledge graph-based methods [16], contextual bandits [17], bandit approaches unifying items and features [18], or topic guided methods [19]. A recent survey [1] provides a good overview of the current status of CRS. Most of the recent works are focused on the underlying methods and algorithms, providing users with a text-only form of interaction. However, multi-turn conversations can be built on any form of interaction or mixed-modality interactions instead of merely textual form [20].

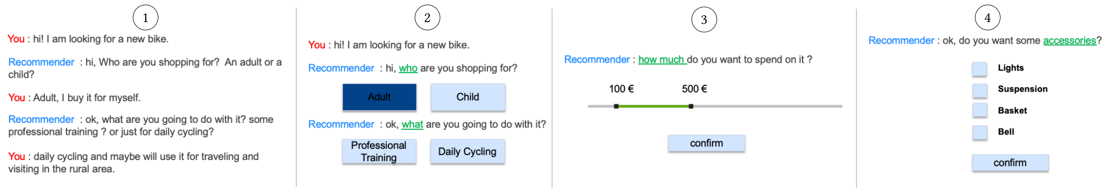
Only limited research has thus far focused on the question how conversational interactions in recommender systems using different modalities impact CRS performance and, in particular, users' perception of a CRS. Several works have investigated different interaction methods, although in other domains. For example, Ciechanowski et al. [21] evaluate different interactions to avoid the uncanny valley effect in chatbots, their interactions including text, voice with human-like avatar animations. In their study, text-based interaction was considered to be more pleasant by the participants, compared to voice interaction with a human-like avatar.

The combination of digital assistants and CRS has been investigated recently, here, the results indicate that a combination of buttons and natural language is particularly beneficial [22]. Jin et al. [23] conducted an experiment to explore the correlation between the user's interaction and personal characteristics. What is interesting for our study is that they deploy several interactive methods in MusicBot: text, voice, button, radio button (ratings). From their results, one can see that participants used buttons most frequently, then radio buttons, followed by text and voice. This indicates that text-only interaction in a CRS might not be the most useful and preferred technique, providing a motivation for our research presented here.

Valério et al. [24] performed a comparison of different chatbot interaction paradigms. Chatbot Kino used only text to communicate with users, while the alternative chatbot Cinemito used text in combination with buttons and images for providing quick feedback. Their analysis revealed that there is not a clearly preferred way of interaction. Their work was a qualitative study ($n = 10$) and mainly focused on user' perception, aiming provide design guidance for chatbots. The study presented in this paper extends existing work by focusing on conversational recommenders, by introducing mixed-modality interaction in CRS and by providing empirical evidence of the benefits of combining interaction modalities for preference capture and critiquing.

3. Mixed-Modality Interaction in CRS

Combining different modalities in human-system interaction can generally bring about various benefits such as increased efficiency of the interaction or better disambiguation in probabilistic input recognition. Text-based or speech-based dialogs provide the user with a natural and flexible interaction style which, however, is also error-prone and often not very transparent since the user needs to anticipate the comprehension capabilities of the system to avoid misinterpretation or rejection of the input. When the input options are limited, selecting from the available options is also mostly quicker, resulting in higher efficiency and often reduced user frustration. While multimodal interfaces may employ a wide range of different modalities [25], we focus in this paper on the visual channel and on the prevalent combination of input techniques in CRS which is textual, language-based dialogue combined with graphical interaction widgets. To distinguish this type of interaction from more general multimodal interfaces, we use the term mixed modality for this combination. Even though restricted in the number of modalities, the design space for interaction based on textual conversation with integrated GUI elements is an under-researched area. To more systematically explore the design options we propose two dimensions for characterizing the interactions.



You : hi! I am looking for a new bike.

Recommender : hi, Who are you shopping for? An adult or a child?

You : Adult, I buy it for myself.

Recommender : ok, what are you going to do with it? some professional training ? or just for daily cycling?

You : daily cycling and maybe will use it for traveling and visiting in the rural area.

You : hi! I am looking for a new bike.

Recommender : hi, who are you shopping for?

Recommender : ok, what are you going to do with it?

Recommender : how much do you want to spend on it ?

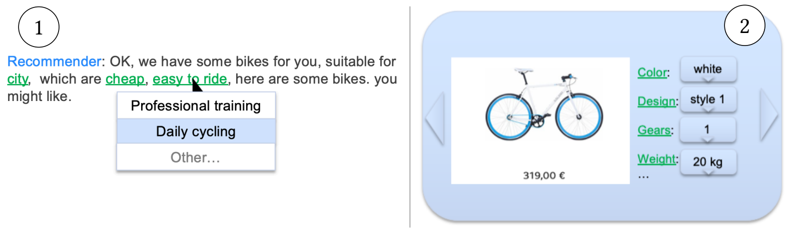
Recommender : ok, do you want some accessories?

□ Lights
□ Suspension
□ Basket
□ Bell

Figure 1: Interaction styles in CRS. In (1), interaction with the virtual agent is exclusively text-based, while the other interaction methods shown allow responses via GUI: In (2) buttons are additionally provided to respond. Range sliders (3) can be used to define continuous values. In case more than one answer option may be submitted, the CRS provides checkboxes.

The first dimension refers to the location where the GUI element is integrated in the conversation flow. We call this the *anchor dimension*. An anchor can be located in or near the user input area where typically buttons are used as shortcuts for otherwise textual user responses. Widgets can also be attached to a presented recommendation itself, be that inline in the textual flow or in a separate recommendation area. Responding to such prompts is essentially equivalent to critiquing the recommendation since the widget will allow the user to change feature values, and thus user preferences, directly on the displayed item. As a third anchor, we propose to embed interactive elements directly in the textual output of the system. This way, the user can respond directly to terms that appear in the output, such as features mentioned or the intended usage of a product. Various options exist for making such feedback available, for instance, through links or embedded drop-down lists. We assume that the user can in all cases also respond by typing a textual question or response thus providing a flexible style of interaction. Depending on where the widget is located it can either serve for specifying preferences in the dialog or for changing, i.e. criticizing, general or item-specific values.

The second dimension refers to the type of interactive element (*widget dimension*) integrated



Recommender: OK, we have some bikes for you, suitable for city, which are cheap, easy to ride, here are some bikes. you might like.

Color: white
Design: style 1
Gears: 1
Weight: 20 kg
...
319,00 €

Figure 2: Critique interactions. With inline critiquing (1) users have the possibility to change features directly within the question/answer of the virtual agent. Modifiable features are highlighted. By clicking on these highlighted words, a drop-down list appears from which other options can be selected. Item-based critiquing (2) allows users to critique the characteristics of features based on a recommended item. Here, the values of the features of the displayed item can likewise be changed by means of a drop-down list.

in the conversation. Here, we consider the standard widgets which can be selected depending on the purpose and constraints of the input. Buttons, checkboxes, drop-down lists or sliders can be offered for responding to system questions, or be attached to a recommended item to show and modify its features. As a novel option in CRS, we propose to also make parts of the system output interactive by embedding links, drop-down lists, or buttons directly in the textual stream to let users react directly on questions, assumptions or suggestions made by the system. In the present study, we investigated six different interaction techniques with respect to their usefulness in CRS (Fig. 1 and 2). In the following section, we will discuss the different forms of providing feedback and critique in more detail.

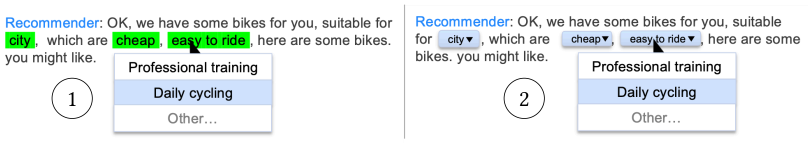
3.1. Critiquing in CRS

Item-based critiquing (Fig. 2 (2)) is a technique that has been gaining considerable interest in recommender systems research. Since recommendations may not meet user preferences, critiquing the features of a recommended item allows users to modify or incrementally refine their preference in an interactive fashion, thus also increasing their control over the recommendations provided [11]. In a CRS, the critiquing approach can be extended to also providing feedback on other concepts that appear in the conversation, such as the system's assumptions about the intended usage of an item, or any other aspect of the user model that is explicitly mentioned in the system output,

Integrating item-based critiquing in CRS could help users conveniently supplement or modify preference when the first recommendation occurs. Besides the limitation of flexibility, another drawback is the learning cost of interaction, users need time to adapt to this interaction.

3.1.1. Inline critiquing

We propose a novel inline critiquing interaction by which users can conveniently state and modify their preferences. The basic idea is that once the system presents a recommendation to the user, it also simultaneously generates a response, summarizing relevant item properties as well as user preferences collected so far. In this summary, some keywords are marked that can be directly modified by the user in the text. This avoids the problem that users would have to refer verbally to previous system responses to criticize their content in a purely text-based interface. This method can, in principle, be applied both to previous system outputs or user inputs. We assume that this form of feedback provides advantages with respect to efficiency as well as error-avoidance. In the following, we describe three different styles used in our study of emphasizing interactive keywords and the widgets used for providing feedback.



Recommender: OK, we have some bikes for you, suitable for city, which are cheap, easy to ride, here are some bikes. you might like.

Recommender: OK, we have some bikes for you, suitable for city▾, which are cheap▾, easy to ride▾, here are some bikes. you might like.

Figure 3: Alternative inline critiquing styles. Alternative display styles to avoid the risk of confusion with hyperlinks. In (1), changeable feature values are emphasized in the text by an eye-catching background color. In 2, the common design pattern for drop-down lists is used to directly illustrate that further options are available.

Underlined Text Fig. 3 (1) shows the first style of critiquing textual elements. The keyword is underlined and shown in a different color, as is common for visualizing web links in HTML. Once the user clicks on this link, a list pops up showing selectable—possibly popular—options. Other feedback can be freely entered in an editable text field. The technique is well-known to most users, although some might misinterpret it as a Web link.

Highlighted Text The second style of indicating interactive keywords uses highlighting, showing the text with a colored background (Fig. 3 (2)). Highlighting can easily attract user's attention and different from underlined text, it does not have hyperlink misunderstanding problem. However, users may interpret it as indicating importance, not interaction.

Drop-down Button Fig. 3 (3) shows the third style of inline critiquing we investigated. This form inserts a button in the text indicating a drop-down list which has equivalent functionality as the other two styles. Buttons are easily recognizable as interactive objects and avoid the problem of potential misinterpretation as is the case with the other two styles, but may look awkward inside a running text.

4. Evaluation

In this section, we describe an empirical comparison of the interaction patterns described in the previous section and their meaningful combination in a mixed-modality CRS against a conventional CRS. We conducted a user study to determine whether there exist preferences for the different techniques and critiquing styles when engaging with conversational agents. In particular, we intended to investigate whether users prefer a purely text-based CRS (TBCRS) or a mixed-modality interaction CRS (MICRS). Besides, we were interested in identifying which interaction modes are favored for different communicative tasks.

4.1. Method

In order to investigate these questions, we performed a study using video prototypes of the different techniques described. To obtain a deeper understanding of users' perception of the techniques, our study was designed to capture quantitative data as well as qualitative feedback from the participants. We split our study into three parts to investigate these research questions.

4.1.1. Comparison between TBCRS and MICRS

Since the focus of this study was centered on obtaining initial insights about users' preferences and the perceived usability of different interaction patterns, we did not yet implement a working CRS with the interaction integrated. Instead, the evaluation was done by means of videos showing the interactions based on a conversational recommender scenario in a fictitious bicycle shop. We created two videos exemplifying different levels of interaction with a fictitious online bicycle CRS. During the first part of the experiment, we presented these two videos to the participants, showing conversations with the CRS in the form of a chatbot. In the videos, a fictitious user tried to find a suitable bicycle for himself by means of the chatbot. Both videos were identical in their content, only the user's interaction possibilities with the chatbot varied as follows:

1. Text-based CRS (*TBCRS*): The conversation with the chatbot is solely text-based. Besides the text-based method, there is no alternative option to respond to the chatbot's questions.[2]

2. Mixed-modality Interaction CRS (*MICRS*): The conversation with the chatbot is both text-based and via direct feedback using i.e. buttons, drop-down lists. Each of the interaction methods described in Section 3 are exhibited. For all actions that the fictive user has to perform in the conventional system through text input, there is an alternative interaction possibility in this version. However, at all times, the user is able to enter simple textual input instead.[3]

Both videos were presented to all participants. Participants were allowed to pause, resume and restart the videos at any time. There was no time limit for watching the videos. We counterbalanced the order of the videos, resulting in a within-subject design.

After each video, participants were asked to fill in a questionnaire. If not indicated otherwise, all questionnaire items had to be answered on positive 1–5 Likert response scales. For this purpose, we asked them to imagine themselves interacting with the chatbot shown and to evaluate the interaction possibilities. To assess user interface satisfaction, we applied the factors of "overall reaction to the software" from the QUIS questionnaire [26], consisting of six items. These items were assessed by means of a polarity profile. Besides, we constructed nine items that

[2] Video of text-based CRS interaction: https://intsys.info/tbcrs

[3] Video of mixed-modality CRS interaction: https://intsys.info/micrs

were specifically intended to evaluate the interaction methods shown. Furthermore, we assessed domain knowledge of participants with self-constructed items and collected demographic data.

4.1.2. Interaction Methods in Detail

During the second part of the study, we sought to obtain more detailed feedback on the six different interaction methods described in Section 3: Free text, buttons, checkboxes, sliders, item-based critiquing and inline critiquing. Here, we successively showed participants the individual interaction methods as screenshots. All interaction methods were already shown in the videos during the first part of the study, thus participants have already seen the interaction process with the respective method. We asked them to rate each interaction opportunity by means of the self constructed questions regarding *enjoyability*, *supportiveness*, *efficiency* and *precision*. Besides, we asked a specific question regarding *critiquing efficiency* for the interaction methods free-text, inline critiquing as well as for the item-based critiquing method. Additionally, we asked what they particularly liked or disliked about each interaction method. This optional questions were open-ended.

4.1.3. Inline critiquing styles

Finally, we aimed to identify the preferred presentation style for the inline critiquing method. Therefore, we asked participants to choose one of three different designs described in Section 4.1.3, as the most appropriate one for directly modifying features in the text. Additionally, they were asked to briefly describe why they selected a particular style. These two question were optional.

4.2. Participants

We recruited 70 participants using Prolific,[4] a tool commonly used for academic surveys [27], of whom 63 finished the study. We pre-selected Prolific users based on the following criteria to maximize quality: (1) participants should be fluent in English; (2) their success rate should be greater than 95 %; and (3) the survey should not be conducted on smartphones or tablets to ensure that the interaction methods shown in the videos and screenshots can be recognized easily. The average duration of the survey was 13.18 minutes (*SD*= 3.28) and each participant received a compensation of 1.25£ if they successfully completed the survey. In our analysis, we only considered participants who watched the videos completely, leaving us with 54 participants.

Demography Out of 54 participants, 32 were female. Their age ranged from 18 to 81 (*M*= 35.2, *SD*= 14.29). The majority had a university degree (46.3 %), 27.8 % had a higher education entrance qualification and 11.1 % had a general certificate of secondary education. The majority originated from the United Kingdom (85.2 %). All other participants originated from South Africa (3.7 %) and further countries (11.4 %). The domain knowledge of participants was rather low (*M*= 2.16, *SD*= 1.12).

[4] https://www.prolific.co

4.3. Results

We present the quantitative and qualitative results of the comparison between the two different video prototypes: TBCRS and MICRS (Section 4.3.1), followed by addressing specific quantitative and qualitative evaluations of each interaction method separately (Section 4.3.2). Finally, we detail the user comments regarding the proposed inline critiquing styles described in Section 4.3.4. Therefore, we will quote exemplary statements made by participants.

4.3.1. Comparison between TBCRS and MICRS

Tab. 1 shows the overall reaction statistics of the two tested CRS, which are derived from participants' ratings of the QUIS questionnaire items and our self-constructed items. To determine whether there are differences between the two conditions, we performed a paired t-test. Unless stated otherwise, preconditions for this and subsequent calculations were met. We used an α-level of $.05$ for all statistical tests.

Table 1
Results from the paired t-test ($df = 53$) between the two conditions. Higher values indicate better results. Values marked with * are significant at a level of $p < .05$. The upper part of the table shows the items from the QUIS questionnaire, whereas the lower part of the table shows the self-constructed items for evaluating the two systems.

	TBCRS		MICRS				
	M	*SD*	*M*	*SD*	*T*	*p*	*d*
terrible / wonderful	3.83	0.99	**3.96**	0.97	-0.880	.383	-0.120
difficult / easy	4.17	0.97	**4.30**	0.92	-0.806	.424	-0.110
frustrating / satisfying	3.50	1.26	**3.89**	1.08	-2.183	**.033***	-0.297
inadequate power / adequate power	3.81	1.20	**4.06**	1.04	-1.390	.170	-0.189
dull / stimulating	3.26	1.31	**3.59**	1.09	-1.685	.098	-0.229
rigid / flexible	**3.63**	1.22	3.50	1.26	0.693	.491	0.094
Messages from the chatbot which prompt for user inputs are clear.	4.11	0.88	**4.20**	0.90	-0.637	.527	-0.087
Learning to interact with the chatbot is easy.	**4.39**	0.69	4.31	0.80	0.704	.485	0.096
I liked the methods for interacting with the chatbot.	3.74	1.15	**3.85**	1.07	-0.685	.496	-0.093
The chatbot gives me the opportunity to react fast and easily to its questions.	4.06	0.94	**4.11**	0.88	-.358	.722	-0.049
It is easy to express which product features I want.	**4.02**	0.94	3.98	0.90	0.265	.792	0.036
With the chatbot I can always easily and efficiently articulate my requirements.	**3.93**	1.01	3.83	1.04	0.552	.583	0.075
It is easy to adjust my preferences.	3.89	1.08	**4.15**	0.90	-1.528	.132	-0.208
It is easy to understand why the chatbot is showing me the recommendations.	4.02	0.42	**4.24**	0.80	-1.806	.077	-0.246
It is easy to criticize the features of the shown recommendations.	**3.54**	1.16	3.35	1.05	1.043	.301	0.142

Except for the "rigid/flexible" item, the MICRS version shown in the video prototype received better average ratings in all elicited items from the QUIS questionnaire. For the item "frustrating/satisfying", we identified a significant difference between the two tested versions. For all other factors tested, we did not observe any significant differences between the two conditions. Furthermore, we could not identify any significant correlations between the values reported here and those reported in the remainder of this work and domain knowledge.

Comparing the comments on each of the video prototypes, it seems that the participants appreciate that it feels "natural to interact" with the virtual agent in the TBCRS version: "I liked this advisor [TBCRS] better as it made me feel like I'm talking to a real person". The virtual agent appears "more human" and "it felt more personal" than the MICRS variant showcased.

Conversely, other participants worry that it might be more difficult to express requirements because "[…] people don't always know exactly what they want and it would be difficult to articulate properties efficiently" and "with free text input, it is difficult to know what answer the chatbot is probing for, which can lead to frustration." In addition to the "potential for higher error margins for misunderstanding the customer," some participants also mentioned difficulties in "[…] reasoning for recommendations because it is less clear what information is considered."

The positive comments on the video prototype of the MICRS mainly refer to the increased efficiency ("It was very efficient and time-saving"; "I like how easy it is to fine-tune my preferences") and possibility of easier specification of personal preferences: "I liked the given options, which saved time and gave ideas you might not have necessarily thought of." Also, the participants perceived the interaction options used as "straightforward and self explanatory."

The reasons why participants disliked this prototype were primarily that its options were rather "specific and seemed less flexible." Also, some participants were not aware that besides the suggested interaction methods, they could continue to provide open text input. Few of them stated that it "did not feel authentic."

Table 2

Results from the paired *t*-test ($df = 53$) for the self constructed items between the free text and GUI based responses (buttons, checkboxes and sliders). Higher values indicate better results. Values marked with * are significant at a level of $p < .05$.

		Free text		GUI-Responses				
Item	Description	*M*	*SD*	*M*	*SD*	*T*	*p*	*d*
Enjoyability	I like this kind of interaction.	3.61	1.32	**4.22**	0.84	-3.234	**.002***	1.388
Supportiveness	This interaction supports me in my search.	3.87	1.14	**4.14**	0.84	-1.702	.095	1.146
Efficiency	This interaction offers me the possibility to articulate my requirements in an easy and efficient way.	3.91	1.14	**4.12**	0.86	-1.299	.200	1.222
Precision	This interaction gives me the opportunity to respond precisely to the chatbot's output.	4.04	0.95	**4.16**	0.76	-0.857	.395	1.059

4.3.2. Free Text vs. GUI-Responses

To compare the different interaction methods, we combined the assessed values for the GUI-based interaction methods (checkboxes, buttons and sliders) into one score. We did this because the text-only approach may be used universally, but not every GUI-based interaction method is equally suitable for all response types.

We conducted a paired *t*-test to compare text-based input with GUI-based methods. As shown in Tab. 2 the GUI-based interaction method for responding to the virtual agent were rated consistently higher. However, we only observed a significant difference for enjoyability.

In addition to the quantitative data, we also analyzed the participants' comments for each of the interaction styles presented. Tab. 4 shows a meta analysis of the comments separated by each interaction style. On average, there were 35.6 positive comments followed by 13.0 neutral comments for the GUI-responses. On average, there were 4.7 negative comments for these interaction styles. All interaction methods received fewer negative comments than the solely text-based interaction. In fact, more than half of the comments on the text-based interaction were neutral or negative (Tab. 4).

Participants commented negatively on the text-only input that “with open-ended responses there are so many ways to respond that I probably would be unsure that my answer would be interpreted correctly.” Others commented positively that “it felt more like writing with a real person” and that they “preferred this kind of input because it is more descriptive.”

Positive comments concerning the GUI-responses often referred to the simplicity (“very clear and helpful”, “straightforward”) and precision of the input: “It is easy to select what you want instead of typing and potentially making a typo which could impact results.” However, other participants noted that these interactions were “potentially limiting” and “narrow”, and that “they may not cover all possible responses.” For other participants, these forms of interaction were “[. . .] too similar to conventional filtering systems in online stores.”

Table 3
Results from the Repeated Measures ANOVA for self constructed items. Higher values indicate better results. Values marked with * are significant at a level of $p < .05$. df_n indicates degrees of freedom numerator. df_d indicates degrees of freedom denominator. For items marked with †, the Greenhouse–Geisser adjustment was used to correct for violations of sphericity.

	Free Text		Inline crit.		Item-based crit.						
Item	*M*	*SD*	*M*	*SD*	*M*	*SD*	df_n	df_d	*F*	*p*	η_p^2
Enjoyability	3.61	1.32	3.98	1.14	**4.28**	1.00	2	106	6.846	**.002***	0.114
Supportiveness	3.87	1.12	3.91	1.15	**4.26**	0.98	2	106	4.140	**.019***	0.072
Efficiency	3.91	1.14	3.89	1.09	**4.28**	1.02	2	106	3.965	**.022***	0.070
Precision	4.04	0.95	3.94	1.12	**4.24**	0.93	2	106	1.892	.156	0.034
Critiquing Efficiency[a] †	3.70	1.18	3.85	1.14	**4.13**	1.12	1.74	91.99	2.534	.092	0.046

[a] *Item description:* This interaction gives me the possibility to criticize the displayed features of the recommendations in an easy and efficient way.

4.3.3. Free Text vs. Critiquing Methods

Additionally, we performed multiple Repeated Measures ANOVAs to compare the text-based interaction with the inline and item-based critiquing variants. The item-based variant was assessed consistently better than the text-only and inline critiquing variants (Tab. 3).

Regarding enjoyability, we found a significant difference between the tested conditions (Tab. 3). Post-hoc analysis revealed a significant difference ($p = .003$) between the interactions free-text and item-based critiquing (-0.667, 95 %–CI[-1.14, -0.19]). Post-hoc tests performed here and in subsequent results were Bonferroni-adjusted.

In terms of *supportiveness*, results indicate a significant difference between the conditions (Tab. 3). Again, post-hoc tests revealed a significant difference ($p = .030$) between the free-text interaction and item-based critiquing (-0.389, 95 %–CI[-0.75, -0.03]).

Furthermore, the *efficiency* of the interaction variants is evaluated significantly differently (Tab. 3). Compared to the free text and the inline critiquing variant, the item-based critiquing is rated better. However, the performed post-hoc tests only revealed a significant difference ($p = .030$) between item-based and inline critiquing (0.389, 95 %–CI[0.028, 0.75]).

Although item-based had higher means for the last two items (*precision* and *critiquing efficiency*) shown in Tab. 3 than inline critiquing and text-based interaction, no significant differences were detected.

Table 4
Summary of the qualitative feedback for the various interaction methods.

		GUI-Responses			Critiquing	
Comments	Free text	Button	Checkbox	Slider	Inline critiquing	Item-based critiquing
# Neutral	17	15	12	12	**18**	12
# Positive	21	35	**36**	**36**	31	35
# Negative	**16**	4	5	5	4	7
Positive	38.9 %	66.8 %			58.5 %	64.8 %
Negative	29.6 %	8.8 %			7.6 %	13.0 %

The received comments regarding the different critiquing methods were also rather positive. Item-based critiquing received more positive comments than inline critiquing. Overall, inline critiquing received the most neutral comments (Tab. 4).

Regarding the item-based critiquing option, participants liked the "[...] ability to directly select different options" and to be able "[...] to criticize options directly based on the given items." Others appreciated that "previous specifications were already taken into account" as well as "[...] being able to specify further features." However, others were critical and noted that "less technological affine users could possibly be overwhelmed."

Some participants felt that inline critiquing was "not as seamless as the other GUI options". In addition, some participants were critical that this interaction option might not be understood by everyone: "If people are not familiar with the Internet (e.g. the older generation) they may not understand how to use this." Others experienced a similar problem in terms of the representation: "I would not have realized they were drop-down lists and assumed they were links [...]."

4.3.4. Inline Critiquing Styles

Next, we present the assessment results of the proposed inline critiquing styles. Of the 54 participants, 50 participants specified a favored style for inline critiquing. From this group, the majority (60.0 %) preferred the *drop-down button* (Fig. 3 (2)). Less often, the other two variants *highlighted text* (24.0 %; Fig. 3 (2)) and *underlined text* (16.0 %; Fig. 2 (1)) were chosen. The other 4 participants did not indicate a preference, but still provided comments on the styles shown.

A χ^2-Goodness-of-Fit-Test shows that there are significant differences between the observed frequencies ($\chi^2(2, N = 50) = 16.485$, $p < .001$). Post-hoc analysis revealed significant differences between *drop-down button* and *underlined text* styles ($p = .003$), and between the *drop-down button* and *highlighted text* ($p = .024$). However, we could not detect any significant difference ($p > .999$) between the *underlined text* and *highlighted text* styles.

Comments from those favoring the *drop-down button* stated that it is "[...] obvious that one can click and alter something [...]" in this variation. Others noted that it is "[...] particularly obvious that there are additional options" and "[...] it cannot be confused with a hyperlink." Other participants stated, that "it appears as an option whereas the other styles might be missed."

Participants who favored the highlighted text style argue that it "[...] stands out most" and it is "easiest" and "clearest to see without no thought, because of the words highlighted with the background color."

Those participants who favored the variant with the underlined text justified their decision by stating that "[...] it looks like a hyperlink and thus makes it clearer that one can click on it."

Others noted that this inline critiquing style “is simple and less cluttered.”

Participants who did not nominate a favorite justified this by stating that all three styles shown are “useful when there are a lot of options to choose from” or that all variations “are clearly illustrating options.”

4.4. Discussion

In this section, we discuss the results of our empirical study. Therefore, we first elaborate on the comparison between the two conditions and subsequently discuss the findings regarding the various interaction methods.

4.4.1. Comparison between TBCRS and MICRS

First, the comparison of the two video prototypes showed slightly better scores for the MICRS condition on most factors, however, using paired t-test, we were only able to observe a significant difference for one factor (frustrating/satisfying). These results are in line with the participants' comments. Here, they rated the MICRS as less error-prone. They also noted that it is simpler to recognize possible options and to specify their preferences. Similarly, the ability to respond quickly to questions is also rated slightly better. As expected, entering feedback through buttons or drop-down lists is faster and easier in comparison to formulating text—presuming that adequate response mechanisms are available respectively that the displayed interaction element is purposeful.

The reason why the MICRS condition is perceived as more stimulating than the comparative condition may also be explained by the observation that participants discover aspects and potential options that they had not considered before. Conversely, the provided options may also lead to the MICRS condition being seen as more rigid. Some participants noted that it would be nice to still be able to submit open text input if, for instance, none of the available options apply. Although this was possible in the prototype video shown, the option may not have been obvious enough. However, in a real system, a text-based input option should remain available to avoid restricting users unnecessarily. This, in turn, can also ensure a more flexible conversation and exploit the potential strengths of an open-ended CRS.

While participants rated the text-only condition in the questionnaires as easier to learn, this partially contradicts the comments provided. Although it may be obvious how to interact with a purely text-based system, it may still be necessary to learn how the virtual agent interprets the input to avoid misunderstandings.

The ratings regarding comprehensibility and interpretability (“Messages from the chatbot which prompt for user inputs are clear”; “It is easy to understand why the chatbot is showing me the recommendations”) are consistent with the provided comments. Here, the MICRS condition is rated better. Due to the highlighted features within the text along with the explicit inputs via i.e. buttons or drop-down lists, it is clearly comprehensible which information the system uses for providing recommendations. Compared to the text-only approach, users receive more visual information throughout the entire conversation. However, we assume that the tendencies observed in our video prototype based analysis could be manifested in an interactive prototype with which participants can interact and thus leverage the features themselves.

4.4.2. Free Text vs. GUI-Responses

While comparing the two systems, we asked the participants to rate the individual forms of interaction shown. Unlike in text-based interaction, an appropriate GUI option must be presented by the system depending on the logical type of response requested. Since the system demonstrated in the video prototype utilized a set of different GUI elements, we aggregated the GUI response options (buttons, checkboxes and sliders) and compared them to the text-based interaction. In the video prototype, we always provided the appropriate GUI response methods.

In terms of enjoyability, the GUI based interaction methods were rated significantly higher, which is in line with the results of the general evaluation. One explanation might be that the use of various input methods is more interesting and thus the interaction is more enjoyable.

Concerning the factor supportiveness, no significant differences were found, although there was a tendency for GUI responses to be rated better. The comments discussed in the previous section support these findings: Non-textual interaction methods seem to support users better, assuming that appropriate options are available.

For the last two factors tested, there were also positive tendencies with regard to the GUI responses. These are rated slightly better than free-text input in terms of efficiency and simplicity, although these differences are not significant. This also corresponds to the participants' comments. As long as the appropriate answer choice can be provided directly, a GUI interaction is considered more efficient since only one click is needed.

Although it is reasonable to assume that appropriately displayed GUI elements would provide more precise feedback, results were comparable. We suspect this is because free text input allows requirements to be expressed that are not presented as options in GUI responses. In case users already have a clear idea of the desired item and which requirements they intend to communicate to the system, they may not require guidance in the form of GUI response options, but can respond more precisely and flexibly with a free, textual interaction.

4.4.3. Free Text vs. Critiquing Methods

When comparing the text-based input with the two critiquing methods, results indicated that the item-based method was rated better than the purely text-based input method in all tested factors. With regard to the factors enjoyability, search support and ease of articulating requirements, the differences between these two interaction methods were significant.

We assume that it is more enjoyable for users to give feedback directly based on specific items than to articulate them in a text. Perhaps when critiquing features of a particular item, the implications of that critique are less ambiguous.

Although the inline critiquing was rated better than the text-only method in some aspects, the differences were rather minor. We suspect that the participants were not entirely aware of how this interaction method was supposed to work. This may be due to the chosen visualization, but also to the fact that this novel interaction method was not sufficiently explained within the video prototypes.

While we did not find any significant differences in the other factors tested, we suspect that the alternatives to text-only interaction may still have advantages. The video prototype method we chose was possibly not capable of identifying them clearly. Here, however, we must take into account the fact that text only feedback may provide more accurate responses if the options provided by the CRS are not what the user expects.

4.4.4. Inline Critiquing Styles

Finally, we discuss the results for the different inline critiquing styles. Although one might assume that all three styles perform similarly, since they all have a prompting character, the participants clearly preferred the drop-down button. We assume that this is mainly due to familiarity with this technique and its clear affordance for changing values.

In contrast, emphasizing active parts of the text by underlining may be confused with a hyperlink—causing users to assume that clicking on it will forward them to another page. This was also reflected in the participants' comments.

Highlighting text with a colored background may not convey clearly enough that it is possible to interact with it and modify options. Rather, users might interpret this highlighting as an indication of importance or as a reference to a help text that appears when hovering the mouse pointer over the highlighted word.

5. Conclusions and Future Work

We investigated a mixed modality interaction approach for CRS and could show in a user study that is positively evaluated by the participants who appreciated the benefits of using diverse interaction techniques within a CRS. Also, the possibility to criticize individual features of the recommended items directly in a CRS as well as the proposed inline text critiquing method was evaluated positively.

Additionally, the non-textual interaction methods were particularly favorably evaluated. Overall, this study suggests that text-only interaction might not be optimal for creating a positive user experience in a CRS. Instead, a combination of different interaction methods is probably preferable. By summarizing and emphasizing relevant item features and user preferences in the text, the explanatory value of CRS responses are probably enhanced, increasing transparency of the system. Enabling users to modify terms directly in the output may also increase the sense of user control. Considering that CRS systems should be accessible from the very first use without detailed instructions, we believe the approach is promising and aim to focus in future work on an easy-to-understand embedding and traceability of changes resulting from applying the inline text and feature critiquing mechanisms.

As a limitation of this work, we are aware that evaluating video prototypes cannot substitute interacting with realistic interactive prototypes. Therefore, we intend to investigate the use of mixed-modality interaction in CRS by implementing a fully interactive prototypes in future work. A particular challenge for building mixed-modality CRS is the question how the interactive options offered to the user can be derived automatically. Potential approaches might be based on leveraging knowledge graph data or information extracted from item descriptions or reviews. Also, suitable response generation techniques are needed that summarize the features the user is likely to criticize in the next interaction step. Furthermore, it will be interesting to explore techniques that can be applied to automatically decide which interaction method is most suitable in a certain conversational context.

References

[1] D. Jannach, A. Manzoor, W. Cai, L. Chen, A survey on conversational recommender systems, ACM Computing Surveys 54 (2021). doi:10.1145/3453154.

[2] Y. Sun, Y. Zhang, Conversational Recommender System, Association for Computing Machinery, New York, NY, USA, 2018, p. 235–244. doi:10.1145/3209978.3210002.

[3] J. Weizenbaum, Eliza—a computer program for the study of natural language communication between man and machine, Commun. ACM 9 (1966) 36–45. doi:10.1145/365153.365168.

[4] K. Ramesh, S. Ravishankaran, A. Joshi, K. Chandrasekaran, A survey of design techniques for conversational agents, 2017, pp. 336–350. doi:10.1007/978-981-10-6544-6_31.

[5] B. Lika, K. Kolomvatsos, S. Hadjiefthymiades, Facing the cold start problem in recommender systems, Expert Syst. Appl. 41 (2014) 2065–2073. doi:10.1016/j.eswa.2013.09.005.

[6] K. Zhou, S.-H. Yang, H. Zha, Functional matrix factorizations for cold-start recommendation, in: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '11, Association for Computing Machinery, New York, NY, USA, 2011, p. 315–324. doi:10.1145/2009916.2009961.

[7] R. Hu, P. Pu, A comparative user study on rating vs. personality quiz based preference elicitation methods, in: Proceedings of the 14th International Conference on Intelligent User Interfaces, IUI '09, Association for Computing Machinery, New York, NY, USA, 2009, p. 367–372. doi:10.1145/1502650.1502702.

[8] W. Wu, L. Chen, Y. Zhao, Personalizing recommendation diversity based on user personality, User Modeling and User-Adapted Interaction 28 (2018) 237–276. doi:10.1007/s11257-018-9205-x.

[9] W. Cai, L. Chen, Predicting user intents and satisfaction with dialogue-based conversational recommendations, in: Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization, UMAP '20, Association for Computing Machinery, New York, NY, USA, 2020, p. 33–42. doi:10.1145/3340631.3394856.

[10] M. F. McTear, S. Allen, L. Clatworthy, N. Ellison, C. Lavelle, H. McCaffery, Integrating flexibility into a structured dialogue model: Some design considerations, in: 6th International Conference on Spoken Language Processing, 2000.

[11] L. Chen, P. Pu, Critiquing-based recommenders: Survey and emerging trends, User Modeling and User-Adapted Interaction 22 (2012) 125–150. doi:10.1007/s11257-011-9108-6.

[12] H. Shimazu, ExpertClerk: Navigating shoppers' buying process with the combination of asking and proposing, in: Proceedings of the 17th International Joint Conference on Artificial Intelligence - Volume 2, IJCAI'01, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2001, p. 1443–1448.

[13] J. Zhang, P. Pu, A comparative study of compound critique generation in conversational recommender systems, in: V. P. Wade, H. Ashman, B. Smyth (Eds.), Adaptive Hypermedia and Adaptive Web-Based Systems, Springer Berlin Heidelberg, 2006, pp. 234–243.

[14] L. Chen, P. Pu, Evaluating critiquing-based recommender agents, in: Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1, AAAI'06, AAAI Press, 2006, p. 157–162.

[15] W. Cai, Y. Jin, L. Chen, Critiquing for music exploration in conversational recommender systems, in: 26th International Conference on Intelligent User Interfaces, IUI '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 480–490. doi:10.1145/3397481.3450657.

[16] K. Zhou, W. X. Zhao, S. Bian, Y. Zhou, J.-R. Wen, J. Yu, Improving Conversational Recommender Systems via Knowledge Graph Based Semantic Fusion, Association for Computing Machinery, New York, NY, USA, 2020, p. 1006–1014. doi:10.1145/3394486.3403143.

[17] X. Zhang, H. Xie, H. Li, J. C. S. Lui, Toward building conversational recommender systems: A contextual bandit approach, CoRR abs/1906.01219 (2019). arXiv:1906.01219.

[18] S. Li, W. Lei, Q. Wu, X. He, P. Jiang, T.-S. Chua, Seamlessly unifying attributes and items: Conversational recommendation for cold-start users, ACM Trans. Inf. Syst. 39 (2021). doi:10.1145/3446427.

[19] K. Zhou, Y. Zhou, W. X. Zhao, X. Wang, J.-R. Wen, Towards topic-guided conversational recommender system, in: Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 4128–4139. doi:10.18653/v1/2020.coling-main.365.

[20] C. Gao, W. Lei, X. He, M. de Rijke, T. Chua, Advances and challenges in conversational recommender systems: A survey, CoRR abs/2101.09459 (2021). arXiv:2101.09459.

[21] L. Ciechanowski, A. Przegalinska, M. Magnuski, P. Gloor, In the shades of the uncanny valley: An experimental study of human–chatbot interaction, Future Generation Computer Systems 92 (2019) 539–548. doi:10.1016/j.future.2018.01.055.

[22] A. Iovine, F. Narducci, G. Semeraro, Conversational recommender systems and natural language: A study through the conveRSE framework, Decision Support Systems 131 (2020) 113250. doi:10.1016/j.dss.2020.113250.

[23] Y. Jin, W. Cai, L. Chen, N. N. Htun, K. Verbert, MusicBot: Evaluating critiquing-based music recommenders with conversational interaction, in: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 951–960. doi:10.1145/3357384.3357923.

[24] F. A. M. Valério, T. G. Guimarães, R. O. Prates, H. Candello, Comparing users' perception of different chatbot interaction paradigms: A case study, in: Proceedings of the 19th Brazilian Symposium on Human Factors in Computing Systems, IHC '20, Association for Computing Machinery, New York, NY, USA, 2020. doi:10.1145/3424953.3426501.

[25] A. Jaimes, N. Sebe, Multimodal human–computer interaction: A survey, Computer Vision and Image Understanding 108 (2007) 116–134. doi:10.1016/j.cviu.2006.10.019, special Issue on Vision for Human-Computer Interaction.

[26] J. P. Chin, V. A. Diehl, K. L. Norman, Development of an instrument measuring user satisfaction of the human-computer interface, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '88, Association for Computing Machinery, New York, NY, USA, 1988, p. 213–218. doi:10.1145/57167.57203.

[27] E. Peer, L. Brandimarte, S. Samat, A. Acquisti, Beyond the turk: Alternative platforms for crowdsourcing behavioral research, Journal of Experimental Social Psychology 70 (2017) 153–163. doi:10.1016/j.jesp.2017.01.006.