Towards Fine-Grained UX Evaluation of Visual CUI Components Using Chatbot Simulations

Patrick Stadler
patrick.stadler@dfki.de
German Research Center for Articifial Intelligence (DFKI)
Berlin, Germany

Stefan Schaffer
stefan.schaffer@dfki.de
German Research Center for Articifial Intelligence (DFKI)
Berlin, Germany

ABSTRACT

HCI researchers are intensifying their work on design guidelines for voice user interfaces (VUIs) and conversational user interfaces (CUIs) and at the same time are trying to incorporate existing knowledge from traditional graphical user interfaces (GUIs). At the current state, however, these guidelines are often still abstract or high-level. We suggest that the CUI community should take a more granular look at these applications and present a small study that evaluated UI components in chatbots using simulations. We found some first leads that users are mostly capable of distinguishing between fine-grained variants of different chatbot UI components. Hence, we suggest researchers, and designers should aim for a more fine-grained perspective when building CUIs. This could also help to establish detailed guidelines for CUI applications.

KEYWORDS

CUI Design; Prototyping; Design Guidelines for CUIs

INTRODUCTION

There are many well established design guidelines for traditional graphical user interfaces (GUIs) and we also see a shifting development into constructing and evaluating such guidelines for conversational user interfaces (CUIs) respectively voice user interfaces (VUIs) as well. However, we currently have not reached a “gold standard” yet (as compared to GUIs). Commercial vendors of CUI respectively VUI applications, such as Google, Meta, etc. provide their own design guidelines for such applications while it is not always clear on which foundations they are based. There are first steps towards CUI

design guidelines, however, they are often pretty high-level; that's why we need to take a closer look at some of these guidelines; In this work we focus on visual user interface components (i.e., widgets) that can occur in CUIs, e.g., a chatbot.

The transition from traditional graphical user interfaces to conversational user interfaces is a common topic of current research in Human Computer Interaction. Investigations and studies on CUIs respectively voice user interfaces can also be relevant for chatbot research, although this can only ever be investigated based on concrete manifestations.

Real chatbots (i.e., actually implemented chatbots) are expensive and complex to produce, although many graphical tools for prototyping exist nowadays. The deployment, changes, feedback - cycle is complex and costly. How would it be if many aspects could also be evaluated with chatbot simulations without requiring subjects to download an app? Especially for small-scale studies, e.g., formulations in direct comparison, the effort is very high, because sometimes different chatbots must be developed and distributed. Direct parallel comparison is usually not possible, and comparisons are done sequentially. With a rapid prototyping platform, such design decisions can be tested in advance with real users from all groups and especially from disadvantaged groups. Simulations can combine the advantages of both worlds: a fast, low-cost prototype and controlled user interaction.

A first small study provides evidence that even laypersons are usually able to detect and evaluate subtle differences of different UI components using a scripted chatbot simulation. As a consequence, we think that a fine-grained evaluation of chatbot features is necessary.

VISUAL CUI COMPONENTS

Modern CUIs, such as chatbots, are not limited to text or voice input respectively output. There is a large collection of visual components that can be used within such applications. These widgets, visual hints, or according to Valério, Guimarães, Prates, & Candello (2017) sign classes, are graphical components of the chatbot user interface that extend the standard text and speech modalities with additional functionalities and modalities. Table 1 is a compilation of found visual UI components in CUI applications. In total, we found 18 such components by literature review (e.g., in Valério, Guimarães, Prates, & Candello (2020); Valério et al. (2017)) or by analyzing available commercial chatbot systems.

Widgets vary in complexity and can range from simple informal text components to complex dashboards with various visual components. However, existing literature has paid limited attention to the usability and acceptance of such widgets in detail, which is why there is space for interesting research questions. As a first step, Ferretti (2022) examined simple widgets, such as date pickers and sliders, in addition to the six visual components established by Valério et al. (2020) in chatbot systems.

Table 1: Overview of found visual components (VCs) in CUI respectively VUI applications

Visual Component	Description
VC 1	Simple, textual response
VC 2	Simple image response
VC 3	Card component
VC 4	Carousel
VC 5	Quick Reply
VC 6	Quick Reply w/ image
VC 7	Persistent menu
VC 8	Media response
VC 9	List / Table
VC 10	Widget w/o interactivity
VC 11	Widget w/ interactivity
VC 12	Map widget
VC 13	Visual Error Correction
VC 14	Date picker
VC 15	Slider
VC 16	Forms
VC 17	"Pay" widget
VC 18	Generic widget

A detailed description and areas of application for each visual component from table 1 can be found in Stadler (2022).

TOWARDS UNIFIED DESIGN GUIDELINES FOR CUIS

Chatbots reveal novel challenges and opportunities for HCI. For that reason, there is a need for design guidelines for applications whose main input and output do not take place via traditional graphical user interfaces (Moore, Arar, Ren, & Szymanski, 2017; Murad, Munteanu, Clark, & Cowan, 2018; Murad et al., 2021), although overlaps exist between traditional GUIs and novel VUIs respectively CUIs. There is also evidence that VUI designers rely on their knowledge and experience from GUI design (Murad & Munteanu, 2022).

> "Although voice interaction is markedly different from graphical user interfaces, classic usability principles are still critical to the quality of the user experience." (Whitenton, 2016)

Murad, Munteanu, Cowan, & Clark (2019) grouped guidelines for both traditional GUI systems and VUI systems based on a literature review. In doing so, they mapped a total of ten guidelines for GUIs against 27 VUI guidelines, combining VUI heuristics from two sources. The presented heuristics are examined or evaluated in places within this work. The idea behind the mapping is the attempt to extend the design principles of GUI systems to VUI systems as well, or to align them, since the research here is already much more advanced (also due to the time spans). The heuristics are generally expressed in abstract terms (for example, heuristic A2: "Make the system status clear") and serve primarily as a basic conceptual tool and are not to be taken as step-by-step instructions for actual design. Nevertheless, they contain many important design principles (heuristic A14: "Use multimodal feedback when available"). Many heuristics for VUIs relate primarily to conversational design, i.e., the dialog manager, and less to frontend design, e.g., the chatbot interface; some can be applied to both use cases (heuristic B7: "Coach a little at a time").

Such design guidelines for chatbot systems also exist from Google, Amazon, and Co., but no publicly available references to corresponding empirical evaluations can be found (Wei & Landay, 2018).

As a first step towards more concrete guidelines, we present a study that examined a fine-grained view on UI components, i.e., details on different variants of widgets in different contexts. These are evaluated within six design questions (refer to table 2) that map combinations of widget variants and contexts.

Table 2: Our six design questions for a first small-scale study

UI Component	Description
D_1	Initiative
D_2	Structured Information
D_3	Visual Error correction
D_4	Quick replies
D_5	Persistent menu
D_6	Carousel

WIDGETEXPLORER CHATBOT SIMULATION

Research shows that it is useful to involve laymen in the design process on Conversational User Interfaces (Sun et al., 2022).

Figure 1: Results for the presentation of the initiative

Figure 2: Results for the presentation of structured information

In this context, design, and creation of inclusive chatbots are time-consuming and expensive in terms of training and adaptation of "real" dialog backends. To be able to evaluate all these aspects, chatbot prototypes must be developed again and again, be it low fidelity in the form of paper prototypes, to elaborate "real" chatbots via Facebook Messenger or other systems (e.g., Rasa), just to be able to evaluate certain aspects of usability. This takes a lot of time, plus there are often technical challenges especially with ML-based chatbots.

WidgetExplorer is our novel rapid prototyping platform for design evaluations of CUIs and chatbots. The ability to display multiple variants directly next to each other opens new possibilities for usability evaluation. In addition, the entire application can be hosted as a service on the Internet, eliminating the need for costly in-place experiments. With WidgetExplorer, up to three variants can be displayed side by side. A more in-depth presentation of WidgetExplorer can be found in Stadler & Schaffer (2023).

STUDY

To evaluate the fine-grained variants described in table 2, we performed a small user study ($n = 33$). Because this paper is intended to only give a first impression of the possibilities, we will limit ourselves to a very brief description of only two exemplarily selected design questions that we present in the next paragraph (Initiative and Visual Error Correction). All other design questions in detail, results, and further information on the study can be found in Stadler (2022).

Initiative

According to Valério et al. (2017), there are two relevant strategies for opening the conversation. In addition, the first interaction steps should be guided or at least simple enough to prevent a new user from getting stuck or lost. This can also be done through UI components, e.g. quick selections, or a carousel. But also a persistent menu can act as part of the initiative. This design question is intended to clarify what the best visual representation respectively UI component for the opening initiative is.

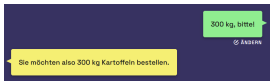
Visual Error Correction

Error recovery is a fundamental challenge in human-machine interaction (Jaber & McMillan, 2020). Errors can occur due to different circumstances, for example due to wrongly classified intents in the dialog manager, or during speech or text input.

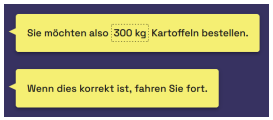
After literature research and analysis of current commercial chatbots, three common types of visual error correction could be identified (see figure 3), which were also implemented and evaluated in this work:

- Error correction via Quick Replies (Bot).

(a) Selection with quick reply



(b) "Change" button



(c) Visual correction using clickable links and pop-ups

Figure 3: Variants of visual error correction

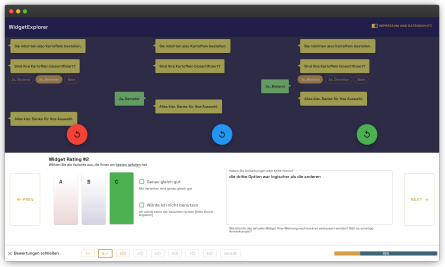


Figure 4: WidgetExplorer during comparison of different representations of a component

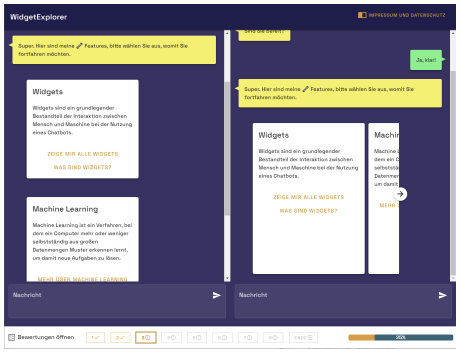


Figure 5: Evaluation of two presentation variants for a carousel widget

- error correction via change button (user), e.g., Solvemate
- Error correction via In-Message popup menu (Bot)

DISCUSSION AND CONCLUSION

To what extent can subtle differences between different chatbot variants be evaluated? This could also be important for aspects of conversation design or UX design, e.g., if certain parameters are to be compared that are difficult to work out in a classic A/B comparison.

Our first small-scale study provides evidence that users are able to recognize subtle differences between different widget variants when they are presented in an appropriate situation.

- A consideration of situation- and task-dependent widgets is worthwhile because it becomes apparent that not every widget is equally suitable in every situation respectively task.
- Prior technical experience or affinity may be a factor in the assessment of widgets but is not necessarily a determining factor. This should be investigated further in a separate study.
- A real dialog system is not necessarily needed to evaluate chatbot components. Simulations and wizard-of-oz setups can also provide realistic data and impressions.

When looking at our research questions, it becomes apparent on the one hand that with the help of the WidgetExplorer, differences can definitely be found, both when looking at the absolute numbers, as well as according to gender-specific distribution and when considering the ATI scores; however, these differences could only be presented, while causes and backgrounds should be explored in further studies, so that actual UI design principles for chatbots can be derived from them. On the other hand, the choice of modality can also depend on the actual situation - this becomes clear in the study on initiative or greeting, among others.

Future studies and work on CUI or VUI applications should take a fine-grained look at different UI components and evaluate them in different situations or tasks to increase the overall quality. Abstract

design guidelines are an important and necessary step, but they are not sufficient and require further clarification.

REFERENCES

Ferretti, B. (2022). *Visual hints on chatbots: E ects on user interaction*. 90.

Jaber, R., & McMillan, D. (2020). Conversational user interfaces on mobile devices: Survey. *Proceedings of the 2nd Conference on Conversational User Interfaces*, 1–11.

Moore, R. J., Arar, R., Ren, G.-J., & Szymanski, M. H. (2017). Conversational UX Design. *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 492–497. Denver Colorado USA: ACM. https://doi.org/10.1145/3027063.3027077

Murad, C., & Munteanu, C. (2022). *"Voice-First Interfaces in a GUI-First Design World": Barriers and Opportunities to Supporting VUI Designers On-the-Job*. 10.

Murad, C., Munteanu, C., Clark, L., & Cowan, B. R. (2018). Design guidelines for hands-free speech interaction. *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct*, 269–276. Barcelona Spain: ACM. https://doi.org/10.1145/3236112.3236149

Murad, C., Munteanu, C., Cowan, B. R., & Clark, L. (2019). Revolution or Evolution? Speech Interaction and HCI Design Guidelines. *IEEE Pervasive Computing*, *18*(2), 33–45. https://doi.org/10.1109/MPRV.2019.2906991

Murad, C., Munteanu, C., Cowan, B. R., Clark, L., Porcheron, M., Candello, H., … Ku, A. (2021). Let's Talk About CUIs: Putting Conversational User Interface Design Into Practice. *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–6. Yokohama Japan: ACM. https://doi.org/10.1145/3411763.3441336

Stadler, P. (2022). *Design und evaluation von multimodalen Interaktionselementen für domänenbasierte Dialogsysteme*. Technical University of Berlin, Germany.

Stadler, P., & Schaffer, S. (2023). WidgetExplorer: A rapid prototyping and evaluation platform to design chatbots for all user groups. *CUI@CHI*.

Sun, L., Liu, Y., Joseph, G., Yu, Z., Zhu, H., & Dow, S. P. (2022). Comparing experts and novices for AI data work: Insights on allocating human intelligence to design a conversational agent. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, *10*, 195–206. https://doi.org/10.1609/hcomp.v10i1.21999

Valério, F. A. M., Guimarães, T. G., Prates, R. O., & Candello, H. (2017). Here's What I Can Do: Chatbots' Strategies to Convey Their Features to Users. *Proceedings of the XVI Brazilian Symposium on Human Factors in Computing Systems*, 1–10. Joinville Brazil: ACM. https://doi.org/10.1145/3160504.3160544

Valério, F. A. M., Guimarães, T. G., Prates, R. O., & Candello, H. (2020). Comparing users' perception of different chatbot interaction paradigms: A case study. *Proceedings of the 19th Brazilian Symposium*

on Human Factors in Computing Systems, 1–10. Diamantina Brazil: ACM. https://doi.org/10.1145/3424953.3426501

Wei, Z., & Landay, J. A. (2018). Evaluating speech-based smart devices using new usability heuristics. *IEEE Pervasive Computing*, *17*(2), 84–96.

Whitenton, K. (2016). *Voice interaction UX: Brave new World...Same old story*.