CHATBOT IN THE MUSEUM - A FIELD STUDY OF USER EXPERIENCE AND MODALITY USAGE

Stefan Schaffer[1], *Eva Schwaetzer*[1], *Aaron Ruß*[1], *Oliver Gustke*[2]

[1]*German Research Center for Artificial Intelligence (DFKI),* [2]*Linon Medien*
stefan.schaffer@dfki.de

Abstract: This paper describes a field study conducted with a museum chatbot at the Städel Museum Frankfurt. The chatbot uses the BERT language model for natural language processing and can be operated via touchscreen as well as via speech input. Prior to the study, hypotheses regarding the user experience of the system were formulated and a system-specific questionnaire was designed, which was used to inquire (among other things) about the perceived quality of the speech output and the frequency of audio guide use in museums. During the interaction with the chatbot, log data was collected and stored in the back-end system. The results show a significant correlation between perceived speech quality and user experience. An exploratory data analysis revealed that participants who used only speech input rated the system as significantly more stimulating than participants who used only touch input. Touch input turned out to be the most efficient input modality in terms of answer correctness and was rated highest regarding pragmatic quality. Interestingly touch input was preferred by younger participants. We discuss our findings and conclude that speech interaction should be seriously considered to create engaging conversational user experiences in museums.

1 Introduction

More and more museums are developing chatbots to assist their visitors and to provide an enhanced visiting experience. Most of these chatbots are developed using chatbot platforms, that provide predefined dialogs [1]. For example, in the museum chatbot *Ping!*, the dialog evolves based on users' decisions between predefined input options[1]. However, such predefined dialogs do not provide a human-like conversation and greatly restrict the range of questions –and answers– that visitors can pose. In recent times, more and more chatbots are appearing that try to employ the latest Natural Language Processing (NLP) techniques for enabling museum-related dialogs and question answering [1]. For example, "The Voice of Art" is an Artificial Intelligence (AI) voice-based interactive guide which allows visitors to freely ask questions about artworks in the Pinacoteca Museum in Brazil [2]. For the museum domain, we can expect increasing use of advanced machine learning techniques based on deep learning for answering freely formulated questions [3], such as language models like BERT [4].

2 Background

The aim of our work is to investigate the perceived user experience (UX) when interacting with a conversational question-answering chatbot in the museum environment. Our system uses BERT to generate answers and Google Cloud services to generate speech output and automatic

[1]https://www.landesmuseum.de/en/ping

 DOI: 10.35096/othr/pub-7075, URN: urn:nbn:de:bvb:898-opus4-70758

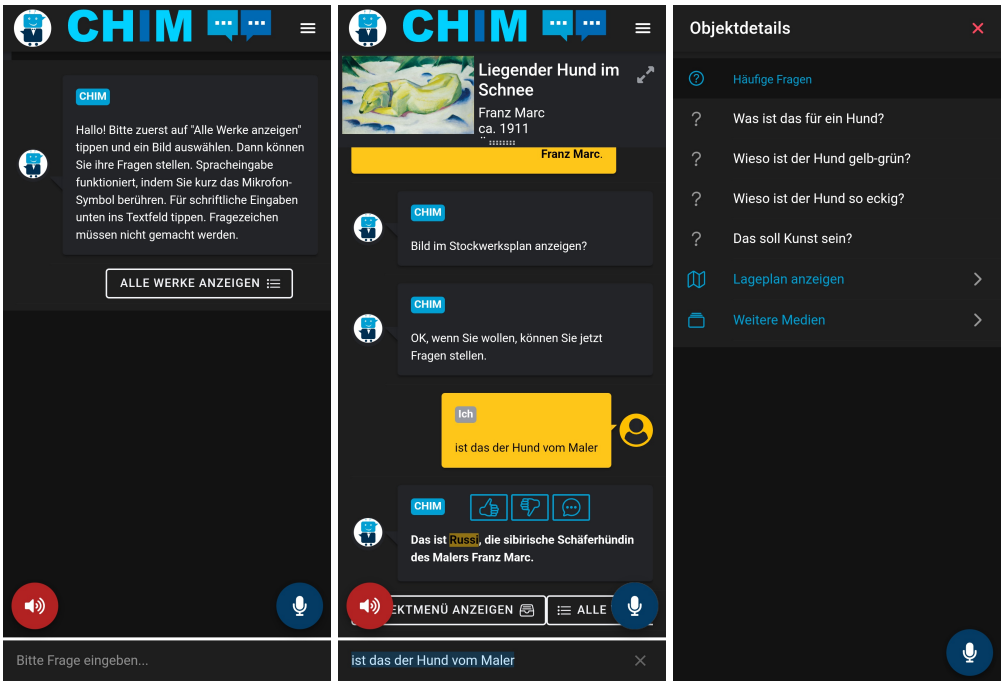

Figure 1 – Screenshots of the chatbot app. Left: greeting message (with active speech output); Center: answer to the question "ist das der Hund vom Maler" ("is that the dog of the painter"; translated answer: "That is Russi the Siberian shepherd dog of the painter Franz Marc"); Right: Displayed "object menu" with a listing of frequently asked questions (FAQs), location map, and other media for the exhibit.

speech recognition. We conducted a field study in the Städel Museum Frankfurt[2] to determine how these technologies affect the perceived UX. In the following we discuss the hypotheses of this paper. System errors are known to have an impact on the UX. However, to our knowledge, the effects of NLP correctness on perceived UX have been poorly studied so far. Our first hypothesis is therefore: [H1:] Participants who received more correct answers also rate the UX of the system as better. The quality of the speech output is an influencing factor for the perceived UX. Since audio guides usually use a lot of high quality speech content that is prerecorded by human speakers, the museum audience is used to an excellent experience in terms of speech quality. We therefore want to evaluate the relevance of the quality of the output of synthesized speech specifically for the museum domain. Our second hypothesis is: [H2:] Participants who rate the speech quality better also rate the UX of the system as better. So far, the museum Städel museum has only a classic audio guide for visitors. Some visitors use audio guides, while others do not. The chatbot represents a new channel for delivering content to visitors. We assume that especially people who know how to use audio guides will have to get used to using the chatbot guide. This adaptation process is an effort. In terms of UX, this additional effort could lead to lower ratings, especially from people who typically tend to use audio guides. Hypothesis three is therefore: [H3:] Participants who tend to use audio guides in museums will rate the UX of the system lower than participants who tend not to use audio guides. In addition, we conducted an exploratory data analysis that focused on observations about the use of input modalities (speech, touch, multimodal). UX, age effects, and the correctness of answers to questions asked during the field test are considered.

3 The ChiM System

The ChiM chatbot was developed as an Android app using the open-source framework Apache Cordova and MMIR framework [5] for speech input and output (for more information see [6]).

[2]https://www.staedelmuseum.de/en

In the chatbot app, users can select exhibits and ask freely formulated questions by voice or text input; the answers can be rated and commented on. For the selected exhibit, the most frequently asked questions about the exhibit, the location map (where the exhibit can be found in the museum), and other media content such as audio files (with audio guide soundtracks) can also be selected via an "object menu" (see Fig. 1).

The basic NLP procedure is based on detecting content types for questions asked about selected artworks from the museum. The approach is based on [2] and an adaptation by [7]. 12 distinct content types can be used to narrow down the answer space as the audio guide texts, the main answer resource, were also annotated using these content types. Important content types are e.g. meaning (questions related to intentions, meanings, or whys, and the stories possibly behind the people and elements depicted in the artwork), author (visitor utterances about the artist's life, which art movement they were part of, or stylistic influences) and content (questions related to what or who is depicted in the artwork, both overall and in detail. Examples: "Is that the baby Jesus on her lap?", "Who are these people?", "Is the dog really sleeping or just pretending?"). The further NLP was implemented as a service with a multi-tiered approach for processing the utterances, using several different models for detecting intents and generating answer-responses (for more details, please refer to our previous publication [6]).

Within the ChiM app, users can select exhibits and freely ask questions by speech or keyboard input. The answers can be rated and commented on (see Fig. 1, center). The previous text input (by touchscreen keyboard as well as by speech) is kept in the input field in a selected state, allowing for corrections by changing the text selection or overwriting them by typing or uttering a new question without changing the selection. For the selected exhibit, an "object menu" provides a list of frequently asked questions about the exhibit, its location map (within the museum), and other media content such as audio files about the exhibits (see Fig. 1, right). Additionally, users can issue speech commands to control the app, for example for giving a rating to an answer (for more information see [6]).

4 Methods

4.1 Preparation and Execution

A concept for the implementation of the field test was developed regarding duration, procedure and methodology used (data collection in line with data protection, standard questionnaires, etc.). A custom questionnaire specially tailored to the project and the standard AttrakDiff questionnaire [8] (on user experience) was implemented in the chatbot app, including a data link to save the collected, anonymized data in a database. In addition, the chatbot app anonymously logs interactions (e.g., selection of a work/image, questions about the work, speech commands) into a database. AttrakDiff measures the attractiveness of the system on 4 scales. It consists of 28 seven-step items whose poles are opposing adjective pairs (e.g., "confusing – clear", "unusual – ordinary", "good – bad"). Each set of adjective items is ordered into a scale of intensity. Each of the averaged values of an item group creates a value for the following scales: scale value for pragmatic quality (PQ), hedonic quality (HQ), subdivided into users' identification with the system (HQI) and into users' stimulation by the system (HQS), and the attractiveness (ATT). The custom questionnaire included questions about the quality of speech and the tendency to use audio guides in museums. These were asked using a 7-point Likert scale. For the field test, 13 smartphones were set up with Android 12 pre-configured with the chatbot app and a kiosk mode app was installed.

The field test was carried out at the Städel Museum from April 26th, 2022, to May 1st, 2022. Two test assistants carried out the field test, instructed the participants, handed out the

devices, and organized the procedure in the museum. The test assistants ensured that the same test schedule was adhered to for all participants. Before starting the visit, the following steps were carried out: 1. welcome; 2. handing out the task description; 3. pseudonym selection by the participants; 4. project description text and data protection settings; 5. assistance in trying out the system; 6. answering any questions that the participants still had. After the visit, the following steps were carried out: 1. if necessary, help with filling in the questionnaires; 2. goodbye; 3. reset system for next participants.

4.2 Participant task and Data preparation

The participants' task was to ask at least one question about at least 6 different objects. If the participants wanted more, they were allowed to ask more questions about as many of the selected objects as they wanted. Participants were allowed to move freely through the museum and had to independently find the objects that are part of the chatbot app. There was no special labelling of the objects in the exhibition. They did not receive any compensation for their participation.

20 participants were excluded from our analyses because they did ask too few questions (less than six) to be able to assess the ChiM system. Furthermore, we excluded some question interactions as system or user errors: (a) In some cases, the exact same question was asked several times in a row. This was probably because the question was displayed (and could be re-sent by pressing a button) until a new question was entered (see sect. 3). Identical questions from the same participants were therefore excluded from the analyses. (b) Also excluded from the analyses were snippets of words whose meaning was not apparent. In total, there were 173 interactions from the included participants that were excluded.

To assess the response quality of the chatbot, all question-answer interactions were annotated by labelling each answer from the chatbot as correct or incorrect. In addition, for correct answers, the relevance was scored on a 5-step scale from 'almost none' to 'perfect'. Preliminary, for the analysis of this paper, the annotation was done by only one annotator. We plan to increase the number of annotations so that each data point is at least annotated by two different persons and publish the data set as part of our future work.

5 Results

5.1 Descriptive Statistics

The sample of valid participations included 108 participants, of which 61 identified themselves as female, 30 as male, one person selected diverse, and 16 persons did not specify. The average age was 34.2 years (SD = 14.9); excluding 34 people who did not specify their age.

A total of 2406 questions were asked about 13 exhibits. Of these questions, 1210 questions were asked by touchscreen-keyboard input (typing) and 1196 questions were asked by speech input. The questions asked by speech input were on average slightly longer (33.6 characters) than the questions asked by keyboard input (26.6 characters). 33 participants (30.6%) used keyboard input only, 22 participants (20.4%) used speech input only, and 53 participants (49.1%) used both modalities. 103 of the 108 participants completed the AttrakDiff questionnaire, and also answered the custom questionnaire for evaluating the chatbot. The participants rated the Attractiveness of ChiM with a mean of 4.47 (SD = 1.00), the Hedonistic Quality-Identity with a mean of 4.07 (SD = 0.94), the Hedonistic Quality-Stimulation with a mean of 4.46 (SD = 1.07) and the Pragmatic Quality with a mean of 4.40 (SD = 0.87). The question about the speech quality was answered by 101 participants and given an average of 4.70 (SD = 1.61) (1 stands for 'terrible', 7 for 'great'). The question about the use of audio guides in the museum was

Table 1 – Correlation coefficients for the three hypotheses.

	Scale	Correlation Coefficient	p-value
Hypothesis 1	ATT	0.01	.89
	HQI	-0.01	.90
(correlation with correct answers)	HQS	0.01	.92
	PQ	-0.07	.51
Hypothesis 2	ATT	**0.50**	.000
	HQI	**0.52**	.000
(correlation with speech quality)	HQS	**0.24**	.02
	PQ	**0.52**	.000
Hypothesis 3	ATT	0.05	.61
	HQI	0.04	.69
(correlation with use of audio guides)	HQS	0.10	.33
	PQ	0.01	.91

answered by 100 participants. The mean value was 4.09 (SD = 1.7) (1 stands for 'under no circumstances', 7 stands for 'whenever possible'). The question about how well the speech recognition worked was answered by 100 participants and given an average of 3.00 (SD = 1.50) (1 stands for excellent, 7 for not at all). The quality of the answers was analysed using the annotated answer-correctness and -relevance described above. Of the total number of questions asked, 61.6% were answered correctly (i.e. annotated as correct). The relevance of the correct answers was also annotated. Of those responses classified as correct, 82.4% were classified as perfect or highly relevant.

5.2 Hypothesis Testing

To test the hypotheses, correlation analyses were performed. The four UX scales of the AttrakDiff questionnaire were each considered individually. To select the appropriate procedure, the collected values were first tested for normal distribution. For this purpose, a Shapiro-Wilk test was performed for each of the four AttrakDiff scales, the evaluation of speech quality, the statement on the use of audio guides and the rate of correct answers. This showed that all variables were normally distributed. Therefore, Pearson's correlation coefficients were calculated. Table 1 shows the correlation coefficients with the associated p-value. Significant results are printed in bold. The computations between the ratio of correct answers and the rating of the chatbot's UX (hypothesis 1) revealed no significant results. The correlation coefficients for hypothesis 2 (speech quality and UX ratings), were all significant, indicating that speech quality is related to the perceived UX of the system at low to medium levels. No significant differences were found for hypothesis 3: For all users as a single group, we could not find a systematic correlation between the frequency of the use of audio guides in museums and perceived UX.

5.3 Discussion

Hypothesis 1 could not be confirmed. We did not find a statically significant correlation between the rate of correct answers and the UX ratings in the four AttrakDiff scales. As most of the correct answers were classified as perfect or highly relevant (81.9%), the relevance of the correct answers is not separately considered in the analysis. With only 61.6% , the number of correctly answered questions is quite low. Qualitative feedback from participants, previously published in [6], suggests a possible explanation for why no effect is evident for the UX. Some participants stated feedback similar to the following: "...if the system doesn't answer my question correctly,

there may still be interesting information about the painting or the artist...". This suggests that some people recognized system errors but did not consider them to be particularly bad. As a consequence, their UX ratings could turn out to be less bad.

Hypothesis 2 could be confirmed. We did find significant correlations between the UX ratings and the perceived speech quality. Regarding all four AttrakDiff scales, the UX ratings were higher if the speech quality was perceived as better. This indicates that the quality of the speech synthesis is an important component when realizing speech-based interactions in the museum context. One possible explanation could be that many museum visitors are used to high-quality speech content from using audio guides. This could also lead to high expectations regarding the speech quality of a chatbot guided tour.

For hypothesis 3, we could not find a significant correlation between the usage frequency of audio guides in museums and the AttrakDiff scales for UX when considering all users as one group. It does not seem to be relevant whether the participants are used to using audio guides. The fact that the functionality of audio guides (in contrast to the functionality of chatbots) is known to many participants does not seem to influence the participants when experiencing the chatbot tour.

6 Exploratory data analysis

6.1 Exploratory Results

We analysed the usage of input modalities by dividing the sample into three groups: individuals who interacted with the ChiM App exclusively via (1) touchscreen-keyboard input (group "touch"), (2) speech input (group "speech"), and (3) multimodal by using both modalities interchangeably (group "multi"). We examined effects of modality usage on the UX ratings of the four AttrakDiff scales. Figure 2 shows the different ratings. Significant differences (marked by *) could be found between the PQ ratings of the speech group and the multimodal group, as well as for HQS ratings of the only speech group and the only touch group. Since these scales are distributed normally, testing for mean differences was performed with a t-test. There were no significant group differences for the other scales.

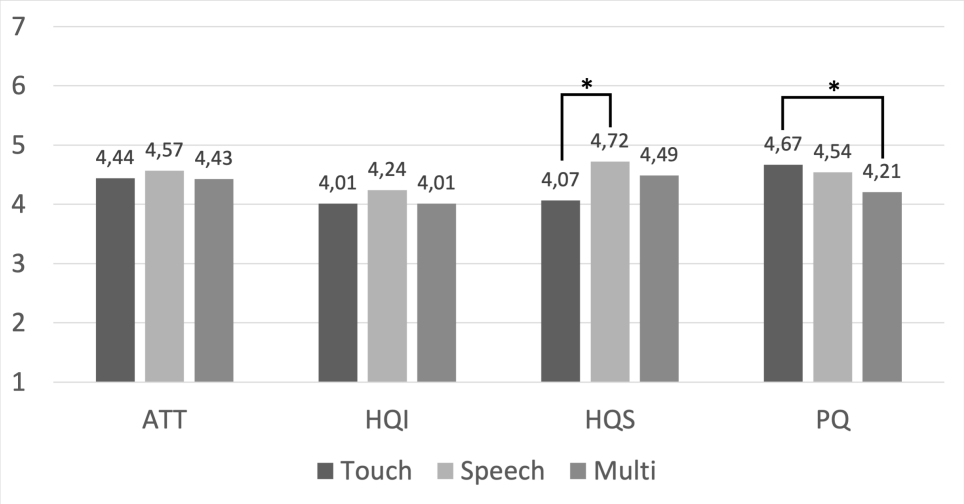

Figure 2 – Means for the UX scales for the different input modalities. * indicates significant differences.

The three groups showed differences in terms of age (compare Fig. 3 left). Individuals who used only speech input (n=22) had a mean age of 37.8 years, individuals who used touchscreen only (n=33) had a mean age of 25.0 years, and individuals who used both modalities (n=53)

had a mean age of 38.3 years. These mean differences were tested for significance using the Kruskal-Wallis test, since age is not normally distributed in all groups. There were significant differences between the group that only used touch input and the group that only used speech input (p = .007), as well as between the group that only used touch input and the group that interacted in a multimodal way (p = .000).

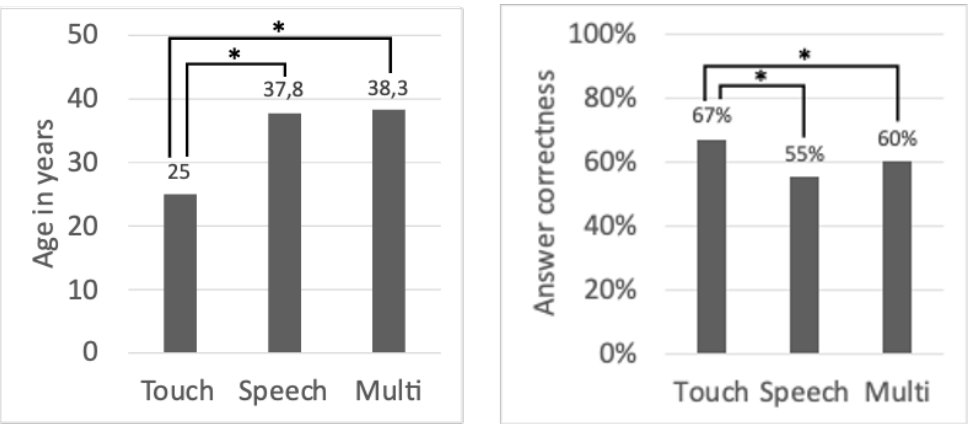

Figure 3 – Left: Mean values for the participants age for the three modality groups. Right: Mean values for answer correctness for the three modality groups. * indicates significant differences.

We also examined whether the mean values of the ratio of correct answers (as annotated) differed for the three groups. Again, the Kruskal-Wallis-Test was used as the variable was not distributed normally in all three groups. The mean differences between the Touch and Speech groups and those between Touch and Multi were significant (with $p_{t-s} = .005$ and $p_{t-m} = .02$; see Fig. 3 right).

6.2 Discussion of Exploratory Results

The analysis of the UX scales for the different input modality groups shows that the group using speech only rated the system significantly more "stimulating" then the group using touch input only. Matching these results, the rating of the multimodal group for HQS lies between that of the touch-only and the speech-only groups. Apparently, the participants found the interaction by speech to be more engaging, creative, original, or challenging. This result can also be considered in combination with the significant correlation between the UX ratings and the perceived speech quality (H2). In terms of PQ, touch-only interaction was found to be significantly better rated than multimodal interaction, while the PQ of speech was somewhere in the middle. We assume that touch inputs were more predictable, since automatic speech recognition errors can also occur with speech inputs. For multimodal use, the system is rated as harder manageable. One possible explanation could be that switching between the modalities is not optimally implemented in the system. Another interesting finding is that participants using speech and multimodal input are on average over ten years older. We pose two possible explanations: 1) Younger people find typing on the smartphone touchscreen easier than older people. 2) Older people were more confident to speak aloud in the museum.

7 Conclusion and Limitations

In the conducted field study, museum visitors were able to have their questions about objects answered by a chatbot during their visit. It could be confirmed that the quality of the speech output decisively contributes to the perceived UX of the system. Participants who rated the speech

output as better also rated the UX of the system as better. Answer correctness and use of audio guide did not correlate with the ratings of perceived UX. Speech turned out to be the most stimulating input modality for participants who only used speech for the asking questions about the exhibits. For conversational museum guides, we therefore recommend that spoken interaction should be seriously considered to create an engaging UX. Younger participants primarily used touch input. Moreover, touch was the most efficient input modality in terms of answer correctness and was rated highest on the pragmatic quality sub-scale. However, whether these results are related must be verified by follow-up studies. The findings are part of the exploratory data analysis and were not formulated as a hypothesis in advance. Therefore, they were not included in the study design.

A clear limitation is that the log data of the study has only been labeled by one annotator so far. Furthermore, the sample is not representative. The participants were visitors of the museum who participated voluntarily. Therefore, the findings cannot be regarded as universally applicable.

References

[1] VARITIMIADIS, S., K. KOTIS, D. PITTOU, and G. KONSTANTAKIS: *Graph-based conversational ai: Towards a distributed and collaborative multi-chatbot approach for museums*. *Applied Sciences*, 11(19), 2021. doi:10.3390/app11199160. URL `https://www.mdpi.com/2076-3417/11/19/9160`.

[2] BARTH, F., H. CANDELLO, P. CAVALIN, and C. PINHANEZ: *Intentions, meanings, and whys: designing content for voice-based conversational museum guides*. In *Proceedings of the 2nd Conference on Conversational User Interfaces*, pp. 1–8. 2020.

[3] GAIA, G., S. BOIANO, and A. BORDA: *Engaging Museum Visitors with AI: The Case of Chatbots*, pp. 309–329. Springer International Publishing, Cham, 2019. doi:doi:10.1007/978-3-319-97457-6_15. URL `https://doi.org/10.1007/978-3-319-97457-6_15`.

[4] KOROTEEV, M. V.: *Bert: A review of applications in natural language processing and understanding*. *arXiv preprint arXiv:2103.11943*, 2021. doi:10.48550/ARXIV.2103.11943. URL `https://arxiv.org/abs/2103.11943`.

[5] RUSS, A.: *Mmir framework: multimodal mobile interaction and rendering*. In M. HORBACH (ed.), *INFORMATIK 2013 – Informatik angepasst an Mensch, Organisation und Umwelt*, pp. 2702–2713. Gesellschaft für Informatik e.V., Bonn, 2013.

[6] SCHAFFER, S., A. RUSS, and O. GUSTKE: *User experience of a conversational user interface in a museum*. In A. L. BROOKS (ed.), *ArtsIT, Interactivity and Game Creation*, pp. 215–223. Springer Nature Switzerland, Cham, 2023.

[7] SCHAFFER, S., A. RUSS, M. L. SASSE, L. SCHUBOTZ, and O. GUSTKE: *Questions and answers: Important steps to let ai chatbots answer questions in the museum*. In M. WÖLFEL, J. BERNHARDT, and S. THIEL (eds.), *ArtsIT, Interactivity and Game Creation*, pp. 346–358. Springer International Publishing, Cham, 2022.

[8] HASSENZAHL, M., M. BURMESTER, and F. KOLLER: *Attrakdiff: Ein fragebogen zur messung wahrgenommener hedonischer und pragmatischer qualität*. In *Mensch & computer 2003*, pp. 187–196. Springer, 2003.