Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 615–619
Marseille, 11–16 May 2020

"Alexa in the wild" – Collecting unconstrained conversations with a modern voice assistant in a public environment

Ingo Siegert
Mobile Dialog Systems, Otto von Guericke University Magdeburg
39106 Magdeburg, Germany
ingo.siegert@ovgu.de

Abstract

Datasets featuring modern voice assistants such as Alexa, Siri, Cortana and others allow an easy study of human-machine interactions. But data collections offering an unconstrained, unscripted public interaction are quite rare. Many studies so far have focused on private usage, short pre-defined task or specific domains. This contribution presents a dataset providing a large amount of unconstrained public interactions with a voice assistant. Up to now around 40 hours of device directed utterances were collected during a science exhibition touring through Germany. The data recording was part of an exhibit that engages visitors to interact with a commercial voice assistant system (Amazon's ALEXA), but did not restrict them to a specific topic. A specifically developed quiz was starting point of the conversation, as the voice assistant was presented to the visitors as a possible joker for the quiz. But the visitors were not forced to solve the quiz with the help of the voice assistant and thus many visitors had an open conversation. The provided dataset – Voice Assistant Conversations in the wild (VACW) – includes the transcripts of both visitors requests and Alexa answers, identified topics and sessions as well as acoustic characteristics automatically extractable from the visitors' audio files.

Keywords: Intelligent Personal Assistants, Digital Personal Assistants, Voice Assistants, Conversational Agent, Smart Speaker, Amazon Echo, Amazon Alexa, public recordings, unconstrained interactions

1. Introduction

In recent years, the market for commercial voice assistants has rapidly grown: e.g. Microsoft Cortana had 133 million active users in 2016 (Osborne, 2016) and the Echo Dot was the best-selling product on all of Amazon's products in the last three holiday seasons (Kinsella, 2018). Furthermore, 72% of people who own a voice-activated speaker say their devices are often used as part of their daily routine (Kleinberg, 2018). Already in 2018 approximately 10% of the internet population used voice control according to (Jeffs, 20178). Mostly the ease of use is responsible for the attractiveness of today's voice assistant systems. Using nothing but speech commands, users can play music, search the web, create to-do lists, shop online, get instant weather forecasts, and control popular smart-home products. In future, the demands and possibilities will increase even more. Voice assistants have become one of the mainstays, with well-known examples like Alexa, Siri, or Cortana from Amazon, Apple, or Microsoft respectively (Osborne, 2016; Kinsella, 2018; Kleinberg, 2018). One can see the prevalence in different areas, such as Smart Home Control, Mobile Assistance, or Operating Systems. One of the main reasons behind this is the given naturalness of speaking as a form of communication in contrast to additional periphery, and on top of that, the independence from additional training based on the single application. Specifically the usage on a mobile phone does not differ from the control of a smart home application. Besides enabling an as simple as possible operation of the technical system, future voice assistants should also allow a natural interaction. A natural interaction is characterized by the understanding of natural actions and the engagement of people into a dialog, while allowing them to interact naturally with each other and the environment. Furthermore users don't need to use additional devices or learn any instruction, as the interaction respects the human perception and experience (Watzlawick et al., 1967; Egorow et al., 2017). Correspondingly, the interaction with such systems is easy and seductive for everyone (cf. (Valli, 2007)). To fulfill these properties, cognitive systems, which are able to perceive their environment and are working on the basis of gathered knowledge and model-based recognition, are needed.

In contrast, today's voice assistant's system functionality is still very limited and not seen as a natural interaction. Especially, when navigating the nuances of human communication, today's voice assistants still have a long way to go. They are still incapable of handling expressions that have semantic similarity but different meanings, are still based on the evaluation of pre-defined keywords/intents, and are still unable to interpret prosodic information as it is needed for an emotional/dispositional understanding (Schuller et al., 2011; Siegert et al., 2015). Furthermore, as it has been discussed in (Porcheron et al., 2018), humans perform a range of conceptual shifts during the interaction with voice assistants. All of these mentioned issues have to be studied and analyzed so that future voice assistants can handle them properly and finally enable a natural interaction. Therefore, research that allows to empirically examine the conversations with voice assistants in an unconstrained and public environment is needed, as still to date little is known about the practical accomplishment of interactions with modern voice assistants. (Nguyen et al., 2016). To do so, the research community demands unrestricted access to such conversational data. To enable investigations on natural language processing, dialogue systems, and computational sociolinguistics for unconstrained interactions with voice assistants, various requirements are set for data recording: a) people talk voluntary, b) people talk unrestricted, c) people talk without fear of being observed/recorded d) people themselves determine beginning and end of the conversation.

To account for the demand of voice assistant interactions in an unconstrained and public environment, a recording platform was designed that engages people to interact with a commercial voice assistant (Amazon's Alexa) and recorded the interaction. The pre-processed data are provided as a corpus, denoted as Voice Assistant Conversations in the wild (VACW). Unrestricted access for research purposes is given to the textual transcriptions, acoustic characteristics and additional meta-data. The corpus itself and first insights are presented in the following submission.

2. Related Work

Available datasets of human-computer interaction (HCI) with modern voice assistants are still quite rare, especially if they are supplemented with acoustics. In the following available datasets utilizing a modern voice interface or reporting about the usage of modern voice interfaces at public places will be discussed shortly.

One dataset studying the use of modern voice interfaces (Amazon Echo) in the everyday life of five households is generated by (Porcheron et al., 2018). During a month-long session six hours of English interactions with the voice assistant could be captured using a special recording device alongside Alexa. Unfortunately, the authors did not specify any number of utterances. Furthermore due to the intimate recordings, the authors just share the textual transcriptions of this data. The advantage of this dataset is the availability of unconstrained conversations due to recordings in the participant's homes as well as the longitudinal study character due to the month-long recording time. Disadvantageous is the relative small number of both participants (five households) and data (six hours). In (Lopatovska and Oropeza, 2018) a data collection of user interactions with Amazon's Alexa in a public academic space is presented. The authors installed a first generation Amazon Echo at a main public hall of the Pratt Institute School of Information for about one month. A total number of 79 sessions comprising 169 interactions could be recorded. Afterwards, the interactions were coded manually into five broad categories (questions, salutations, requests, unclear/inappropriate, and commands). The authors stated that most of the interactions are from the questions category with a focus on the functionality of Alexa, factual questions and weather information.

Another dataset providing recordings of interactions with Amazon's Alexa is the Voice Assistant Conversation Corpus (VACC) (Siegert et al., 2018). Its is a collection of spoken German conversations between a user, a confederate, and Amazon's Echo Dot. The recordings took place in a living room-like surrounding so that the participants could get into a more informal communicational atmosphere compared to a laboratory setting. During each experiment, a user was solving various tasks with Alexa, e.g., making appointments or answering questions of a quiz. While solving the tasks, the user was in some cases cooperating with a confederate, e.g., discussing possible appointment dates or strategies to ask Alexa for the quiz questions. The confederate was only assisting the user and has never talked to Alexa directly. This dataset comprises 17h of recordings from 27 participants with a total number of approx. 3 000 device-directed and approx. 1400 human-directed utterances. Although this dataset comprises much more data then the previous ones, the conversations are neither unconstrained nor recorded in a public environment, as the participants where aware of the experiment, the recording and the study.

In September 2019 Google released two open-source datasets related to the study of voice assistant interactions, the Coached Conversational Preference Elicitation (CCPE) (Radlinski et al., 2019) and Taskmaster-1 (Byrne et al., 2019). Both datasets consist of dialogs between a human partner and a Wizard-of-Oz (WOZ)-ed system. CCPE includes 500 dialogs (12k utterances) about movie preferences. The Taskmaster-1 has a total number of over 13k dialogs from six different categories: ordering pizza, creating auto repair appointments, setting up ride service, ordering movie tickets, ordering coffee drinks, and making restaurant reservations. Although both datasets are unscripted and quite huge, they do not depict a public unconstrained interaction, as at least the topic was pre-defined and the participants were aware of the conducted study.

For the sake of completeness, it should also be mentioned that the Amazon researchers itself have a non-public dataset of 250 hours (350k utterances) of natural human interactions with a voice controlled far-field device cf. (Mallidi et al., 2018) used for their research to improve Alexa devices. Unfortunately no further information is given about this dataset.

Besides these datasets, some studies reporting about the use of voice assistants in public places are available. Although they do not supply a dataset, these studies give valuable hints about public studies with voice assistants.

One public interaction experiment utilizing an Amazon Echo was made 2016 at an art museum in Philadelphia (Barnes Foundation) (Bernstein, 2016). The authors stated that the visitors were not aware of the technology or sometimes were to shy and therefore did just rarely interact with the voice assistant. The visitors who chose to interact with Alexa experienced difficulties due to the system's inability to "understand" the name of foreign artists and thus the interaction was canceled quite fast. The same observation was made with an Alexa exhibit at the Museum of Modern Art (MoMA)in New York City. Again the language understanding ability was a major barrier in the interaction with the voice assistant (Moore and Pan, 2017).

Overall, the reported datasets and investigations are just a selection, but they illustrate the main difficulties in recording unconstrained interactions with voice assistants. These difficulties can be found in enormous effort, which has to be made to record a sufficient number of interactions – one moth does just give a small number of interactions. Furthermore, just installing an interaction device at a public place does not guarantee interactions at all, due to the shyness of people interacting with voice assistants and the lack of functionality, which quickly leads to a termination of the interaction. To overcome these issues, an interaction was designed allowing participants to discover Alexa's capabilities on their own while using the lack of functionality explicitly as a show-case of the interaction. Due to a gamification character participants are engaged to interact freely with Alexa without a force to use the assistant to reduce the shyness.

3. The Setting of Public Voice Assistant Interaction Recordings

Figure 1: Picture of the 2019 exhibition at the MS Wissenschaft featuring AI topics (Copyright: Ilja Hendel/WiD).

The recording of these public and unconstrained voice assistant interactions took place during the science exhibition of the "MS Wissenschaft" from May until October 2019. The "MS Wissenschaft" is a yearly touring exhibition with a distinct topic, The topic of 2019 was artificial intelligence (AI)[1]. 27 different exibits all of them related to artificial intelligence could be visited on the ship. The exhibition is placed onto a ship which travels through different cities in Germany and Austria. In total 31 cities were visited, at each station the ships stays for 3 to 5 days. The exhibition is primarily aimed at school classes but also at interested adults, see Fig. 1. It is supported by various event formats (lectures, meet the scientists, film screenings and panel discussions) to attract visitors. More than 85 000 people with more than 500 classes visited MS Wissenschaft.

Figure 2: Picture of the exhibit used to record interactions with modern voice assistants.

One of the exhibits was developed by us and used to show the lacking functionality of today's intelligent voice assistants for a specific task. Simultaneously, this exhibit was used to record the interaction data. Fig. 2 shows the utilized exhibit. The story of the exhibit was to use modern voice assistants as joker within a knowledge quiz. Visitors had to answer random questions presented on a screen, previously the visitor could set the difficulty level (child, teenager, adult). Possible answers were given and the visitor could use the voice assistant to get the correct answer. Thereby, the questions were designed in such a way that Alexa is not able to give the answer directly. Thus, the visitor has to ask for partial steps or reformulate the question. Furthermore, a time constraint of 30 sec is used, which can be extended by the visitor after expiration. In case of inactivity, the actual question is exited and the start screen is shown so that the next visitor can interact with the exhibit. To increase the attention and enable group interactions, the output of the voice assistant is played back via loud speakers. The interaction presented on the screen and the interaction conduced with Alexa were not connected to each other, so that the visitors had to give the details about the question to Alexa. The exhibit was accompanied with explanations stating that actual voice assistants has a lack of functionality in helping users in answering questions. We assume that due to the exhibition character and the gamification character the visitors shyness to interact with a voice assistant was reduced,In addition, the exhibition explicitly invited visitors to try and play around with the exhibits, by which it was hoped to get a lot of interaction data.

This exhibit used an unmodified version of the Amazon Echo Input. It was denied to use a specially designed skill, as the exhibit should be used without instructions or supervision. For the recording the Amazon Echo Input was directly used. Therefore only the human input is available as audio record. The output of the voice assistant is just available as transcription. The data-collection is not finished yet, as the use of the exhibit at further events is already planned.

4. Dataset overview

Duration	39.9h
# Utterances	32 758 (37 563)
# Sessions	7 144
Language	German
Annotation	transcriptions, topics

Table 1: Key characteristics of the VACW dataset.

The exhibition lasted 126 days and comprises a total of 37 563 items, from which are 32 758 device directed utterances with a total duration of about 40h. The remaining number of items are failures in the activation of Alexa, either due to background noise (11,49%) or phonetically similar words (3,18%). The recording had with full approval of the data security officer of the University.

For each speech utterance the logging data is stored, as well. This data included the timestamp of the interaction and the transcription of the user query (the way it was transcribed/understood by Alexa) as well as Alexa's response, if applicable. The duration of each utterance was 5.21 sec on average, with a minimum duration of 0.1 sec and a maximum duration of 31.2 sec. Further details about the dataset can be found in Table 1.

In order to preserve the anonymity of the visitors and do not raise the awareness of being part of a research study, it

[1] Visit https://ms-wissenschaft.de/ for more information about this science exhibition.

was not asked for additional information such as age, gender or other sociodemographic data (personality, experience with technical systems). But, based on the time-stamps of the interactions the visited cities were identified, to allow analyses of regional differences.
Similarly to (Lopatovska and Oropeza, 2018), an interaction is definded as a single utterance from a visitor and if available the corresponding response from Alexa. In the same manner a session is defined as an ongoing conversation with Alexa, which is characterized by the fact that there is only a small time-span between consecutive user requests. A statistical content analysis of the interactions and timestamps was used to determine the number of sessions and interactions with Alexa. Furthermore the transcripts are used to determine the topics of the interactions.
To give an overview which types of interactions are present in the dataset, the following topics as an adaptation of the ones used in (Lopatovska and Oropeza, 2018): were defined quiz-related questions and other factual questions, Alexa features, movie/tv, music, weather, time/date, playing around, salutations, games, recommendations, and inappropriate see Table 2 for examples and percentage distribution. Due to the design of the exhibit, it was expected that most of the requests are related to the presented quiz. But visitors also asked questions not related to the quiz, e.q. "Who are Romulus and Remus?". Similarly to (Lopatovska and Oropeza, 2018), Alexa features and time/date related requests are occurring quite often. The topics salutations, games, movie/tv, and recommendations are uttered comparable seldom, whereas inappropriate request (swear words, insults) are occurring quite often, this even includes a few racist expressions.

Topics	Frequency
Quiz-Questions	41.3%
Other-Questions	10.1%
Alexa features	16.0%
Time/Date	7.4%
Music	5.6%
Playing around	3.2%
Weather	1.4%
Inappropriate	1.4%
Saluations	0.8%
Games	0.4%
Movie/TV	0.2%
Recommendations	0.1%
Other	12.1%

Table 2: Types of visitor interactions with Alexa during the exhibition, with examples and frequency.

The *Alexa features* are mostly related to volume changes ("louder"), stop commands ("stop") or questions to Alexa ("How old are you"). The topic *time/date* contains mostly utterances asking for the actual time ("What time is it?") or date ("What day are we having today?"). Requests related to *music* are mostly requests to play a specific song, songs from an artist, or start a radio station. The category *playing around* mostly covers request to tell a joke, fill the shopping list, or set reminders. Requests related to *weather* all asked for the actual weather, temperature at a specific place, mostly the actual place of the exhibition, sometimes for other (popular) cities. For *salutations*, mostly the request is initiated with a "Hello", sometimes a "Goodbye" is used. A few visitors also explained why he/she has to leave: "We have to go home now we still want to have a barbecue"[2]. In *games*, visitors asked to play a specific game or asked for information about game consoles. The topic *movie/tv* is characterized by request to open youtube, or films with a specific actor. *Recommendations* contains requests about actual festivals and events at a specific place. The large amount of request summarized as *other* either contain requests that would fit to multiple topics, as the request contains several statements in once (35%), just contain the activation word "Alexa" (30%), are general one-word requests (20%), or the topic could not be identified, due to transcription errors (15%).
To roughly estimate the number of sessions (ongoing interactions with Alexa), the timestamps of the requests were used. A threshold of 30 sec was defined as boundary between different sessions. This boundary was determined on the basis of previous interaction experiment with Alexa where the maximum length of a system answer was 24 sec and the maximum delay between consecutive requests was less then 15 sec (Siegert et al., 2018). Using this threshold, 7 144 sessions could be identified within the VACW dataset. It has to be noted that these sessions contain both single-user HCI as well has multi-party HCI, as the system output was played back via loudspeakers and thus groups of visitors are implicitly encouraged to talk to Alexa.
Other important observations made, are the sometimes used politeness terms ("thanks", please"), as well as the mention of other voice assistants, either as activation word ("Hey Siri Alexa") or in connection with a question ("Do you know Cortana? ") or a remark ("Siri is better than you"), sometimes the term google is used to initiate a web search ("Alexa google the age of XY"). The also possible alternatives "echo" (6 times) and "computer" (0 times) to activate Alexa are used by just a few visitors. Table 3 indicates the usage distribution of different activation words.

Activation word	Occurrences
Alexa	8 732
Alexa (multiple times)	314
Hey Alexa	16
Hi Alexa	1
Hey Siri	3
Hey Google	3
Google	1

Table 3: Distributions of different "activation" words. As Alexa sometimes allow to utter follow-up requests, not all utterances need an activation word and therefore this number is smaller than the total number of utterances.

While previous findings have already been described in other papers (Porcheron et al., 2018; Bernstein, 2016), VACW also allows two further observations. Some visitors – presumably as a joke – asked if Alexa liked them, wanted to be his/her friend, or if Alexa would merry them (50 occurrences). A

[2]Original sentences are in German, translation is done by the author.

further observation is that a few visitors expressed their "fear for surveillance" asking Alexa if they records them "even if it is switched off" or if "the data is send to a intelligence service" (40 occurrences). Furthermore, the presence of the voice assistant's responses allows to analyze which user requests were successfully answered, which made problems and especially which strategies the user used to continue the dialog, if the problem could be solved or if the user resigns. This issue has not been analysed, yet. But the data contains around 15% of user requests that are either not answered or answered with a general message that Alexa did not understand the request. This shows that this data contain valuable aspects to analyze this issue further.

5. Conclusion and Outlook

In this paper, a new dataset on natural single- and multi-user HCI recorded in an public environment is presented. The focus of this dataset is on unconstrained interactions with a commercial voice assistant system. The recordings are pursued during a science exhibition lasting over 126 days. In future it is planned to use this exhibit in other science exhibitions to even collect further interactions.

As the participants were not directly aware of the recording and the interaction was gamified, a wide range of interactions could be observed. First insights of the various occurring topics as well as specific highlights (richness in activation words, surveillance related questions, or marriage requests) have been highlighted. Also the availability of Alexa's responses offers many possibilities to investigate the HCI strategies. Furthermore, as the visitors' audio stream was recorded as well, various acoustic characteristics can be provided for further analyses. The presented dataset is one of the very few datasets providing a huge number of utterances (30k) allowing a broad and comprehensive analysis of interactions with voice assistants. Some examples for analyses are the reasons why politeness terms are used, the difference between single and group interactions, regional differences in the usage of voice assistants, and the strategies in formulating requests.

6. Availablility

The Voice Assistant Conversations in the wild (VACW) is available for collaborative research related to interactive dialog systems including data management (information extraction and retrieval, dialog design, and speech transcription), conversational user interfaces (acceptance, natural language processing, user groups), assistive technologies upon request.

7. Bibliographical References

Bernstein, S. (2016). What can alexa teach us about the barnes? [blogpost]. medium.com, December. [Online; posted 06-Dec-2016].

Byrne, B., Krishnamoorthi, K., Sankar, C., Neelakantan, A., Duckworth, D., Yavuz, S., Goodrich, B., Dubey, A., Kim, K.-Y., and Cedilnik, A. (2019). Taskmaster-1: Toward a realistic and diverse dialog dataset.

Egorow, O., Siegert, I., and Wendemuth, A. (2017). Prediction of user satisfaction in naturalistic human-computer interaction. *Kognitive Systeme*, 1.

Jeffs, M. (20178). Ok google, siri, alexa, cortana; can you tell me some stats on voice search? The Editr Blog, January. [Online; posted 8th-Jan-2018].

Kinsella, B. (2018). Amazon echo device sales break new records, alexa tops free app downloads for ios and android, and alexa down in europe on christmas morning. voicebot.ai, December. [Online; posted 26-Dec-2018].

Kleinberg, S. (2018). 5 ways voice assistance is shaping consumer behavior. think with Google, January. [Online; posted Jan-2018].

Lopatovska, I. and Oropeza, H. (2018). User interactions with "alexa" in public academic space. *Proceedings of the Association for Information Science and Technology*, 55(1):309–318.

Mallidi, S. H., Maas, R., Goehner, K., Rastrow, A., Matsoukas, S., and Hoffmeister, B. (2018). Device-directed utterance detection. In *Proc. of the INTERSPEECH'18*, pages 1225–1228.

Moore, S. and Pan, D. (2017). A case study on using voice technology to assist the museum visitor. In *MW17: Museums and the Web 2017*.

Nguyen, D., Doğruöz, A. S., Rosé, C. P., and de Jong, F. (2016). Computational sociolinguistics: A survey. *Computational Sociolinguistics*, 42(3):537–593.

Osborne, J. (2016). Why 100 million monthly cortana users on windows 10 is a big deal. TechRadar, July. [Online; posted 20-July-2016].

Porcheron, M., Fischer, J. E., Reeves, S., and Sharples, S. (2018). Voice interfaces in everyday life. In *Proc. of the 2018 ACM Conference on Human Factors in Computing Systems (CHI '18*. ACM, April.

Radlinski, F., Balog, K., Byrne, B., and Krishnamoorthi, K. (2019). Coached conversational preference elicitation: A case study in understanding movie preferences. In *Proceedings of the Annual SIGdial Meeting on Discourse and Dialogue*.

Schuller, B., Batliner, A., Steidl, S., and Seppi, D. (2011). Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Commun*, 53:1062–1087, 11.

Siegert, I., Böck, R., Vlasenko, B., Ohnemus, K., and Wendemuth, A. (2015). Overlapping speech, utterance duration and affective content in hhi and hci - an comparison. In *Proc. of 6th IEEE Conference on Cognitive Infocommunications (CogInfoCom 2015)*, pages 83–88, Györ, Hungary, Oktober.

Siegert, I., Krüger, J., Egorow, O., Nietzold, J., Heinemann, R., and Lotz, A. (2018). Voice Assistant Conversation Corpus (VACC): A Multi-Scenario Dataset for Addressee Detection in Human-Computer-Interaction using Amazon's ALEXA. In *Proc. of the 11th LREC*, Paris, France, may.

Valli, A. (2007). Notes on natural interaction. Technical report, University of Florence, Italy, 09.

Watzlawick, P., Beavin, J. H., and Jackson, D. D. (1967). *Pragmatics of Human Communication: A Study of Interactional Patterns, Pathologies, and Paradoxes*. Norton, Bern, Switzerland.