frontiers in Computer Science

METHODS
published: 21 June 2021
doi: 10.3389/fcomp.2021.686050

Crowdsourcing Ecologically-Valid Dialogue Data for German

Yannick Frommherz[1]* *and Alessandra Zarcone*[2]

[1]Audio and Media Technologies, Semantic Audio Processing, Fraunhofer Institute for Integrated Circuits IIS, Erlangen, Germany, [2]Audio and Media Technologies, HumAIn, Fraunhofer Institute for Integrated Circuits IIS, Erlangen, Germany

Despite their increasing success, user interactions with smart speech assistants (SAs) are still very limited compared to human-human dialogue. One way to make SA interactions more natural is to train the underlying natural language processing modules on data which reflects how humans would talk to a SA if it was capable of understanding and producing natural dialogue given a specific task. Such data can be collected applying a Wizard-of-Oz approach (WOz), where user and system side are played by humans. WOz allows researchers to simulate human-machine interaction while benefitting from the fact that all participants are human and thus dialogue-competent. More recent approaches have leveraged simple templates specifying a dialogue scenario for crowdsourcing large-scale datasets. Template-based collection efforts, however, come at the cost of data diversity and naturalness. We present a method to crowdsource dialogue data for the SA domain in the WOz framework, which aims at limiting researcher-induced bias in the data while still allowing for a low-resource, scalable data collection. Our method can also be applied to languages other than English (in our case German), for which fewer crowd-workers may be available. We collected data asynchronously, relying only on existing functionalities of Amazon Mechanical Turk, by formulating the task as a dialogue continuation task. Coherence in dialogues is ensured, as crowd-workers always read the dialogue history, and as a unifying scenario is provided for each dialogue. In order to limit bias in the data, rather than using template-based scenarios, we handcrafted situated scenarios which aimed at not pre-*script*-ing the task into every single detail and not priming the participants' lexical choices. Our scenarios cued people's knowledge of common situations and entities relevant for our task, without directly mentioning them, but relying on vague language and circumlocutions. We compare our data (which we publish as the CROWDSS corpus; n = 113 dialogues) with data from MultiWOZ, showing that our scenario approach led to considerably less scripting and priming and thus more ecologically-valid dialogue data. This suggests that small investments in the collection setup can go a long way in improving data quality, even in a low-resource setup.

Keywords: dialogue data, voice assistants, crowdsourcing, Wizard-of-Oz, German, ecological validity, situated knowledge

OPEN ACCESS

Edited by:
Jessica Martina Szczuka, University of Duisburg-Essen, Germany

Reviewed by:
Marius Hamacher, University of Duisburg-Essen, Germany
Christos Troussas, University of West Attica, Greece
Sebastian Kahl, Bielefeld University, Germany

***Correspondence:**
Yannick Frommherz
yannick.frommherz@iis.fraunhofer.de

Specialty section:
This article was submitted to Human-Media Interaction, a section of the journal Frontiers in Computer Science

Received: *26 March 2021*
Accepted: *26 May 2021*
Published: *21 June 2021*

Citation:
Frommherz Y and Zarcone A (2021) Crowdsourcing Ecologically-Valid Dialogue Data for German. Front. Comput. Sci. 3:686050. doi: 10.3389/fcomp.2021.686050

INTRODUCTION

Recently, smart speech assistants (SAs) have found their way into the lives of more and more people (Byrne et al., 2019; Yuan et al., 2020). Despite their increasing success, their main applications are currently limited to command-and-control via short commands ("Play the next song"), to simple question-answering or to performing tasks via *slot filling*, that is a conversation pattern which allows the SA to request pieces of information from the user in a structured manner (Asri et al., 2017). Complex tasks such as *booking a table at a restaurant* are typically simplified as instances of slot filling (i.e., the interaction is shaped by the SA requesting slots like the *time* or *number of people* for the booking, and the user informing the SA about them). However, SAs could potentially help humans solve complex tasks in a more sophisticated way, for example, by building common ground, negotiating information or supporting deviations from the "happy path". For this, single-turn or strictly-designed interfaces are not adequate anymore, but a flexible and efficient way of interacting over multiple turns is required, not unlikely the way humans interact in a dialogue while they collaborate on a task.

Dialogue, however, comes with a new set of challenges for machines, including, but not limited to, context-sensitivity, anaphora, ellipsis and dynamic error management (Williams and Young, 2007; Grosz, 2018; Serban et al., 2018; Byrne et al., 2019; de Vries et al., 2020). SAs need to be able to handle these dialogue-specific phenomena, not only to assist in complex tasks, but also to make SA interactions more natural in general. In fact, studies suggest that humans generally prefer dialogue as a mode of interacting with SAs (de Vries et al., 2020).

One way to model dialogue in SAs and make SA interactions overall more natural, is to train the underlying natural language processing (NLP) modules on linguistic data that is representative of natural dialogue (Rieser and Lemon, 2011). To collect large-scale datasets, recent approaches have leveraged scenarios generated from simple templates where entity placeholders are replaced by possible entity surface forms (e.g., "Find a [CUISINE] restaurant" > "Find a Japanese restaurant"; e.g., Budzianowski et al., 2018; Wang et al., 2012). However, such template-based collection efforts may provide too fixed of a script for the dialogue and prime the participants into using specific words (de Vries et al., 2020).

Therefore, in this paper we present a framework to crowdsource dialogue data for the SA domain which is specifically aimed at limiting the kind of bias induced by the template-based approach while still allowing for collecting dialogue data in a low-resource and scalable way. Crowdsourcing, that is, relying on a large, remotely-located pool of workers who perform small tasks on a dedicated website like Amazon Mechanical Turk (AMT), is now a well-established and reliable method for collecting large-scale datasets in a time- and cost-effective way (Schnoebelen and Kuperman, 2010; Buhrmester et al., 2011; Garcia et al., 2020). More concretely, our approach pursues three goals: 1) The collected data should be of good quality, that is, the dialogues should be coherent, diverse and natural. 2) The data should be collected in a low-resource fashion, that is, with as little overhead as possible and relying to the greatest extent possible on existing technologies. 3) Our approach should allow for collecting data in languages other than English (in our case German), where fewer crowd-workers may be available.

Our contribution with this paper is threefold. 1) We describe a novel approach for collecting dialogue data which strikes a balance between limiting researcher-induced bias and a low-resource, scalable data collection setup. Rather than using template-based scenarios for eliciting the dialogues, we used *situated scenarios* formulated in a way that was aimed at reducing bias (scripting and priming). Our scenarios were designed to tap into the participants' situated knowledge, in order to afford them the opportunity to go about solving their task more freely, while at the same time avoiding explicit reference to relevant entities. Additionally, as method sections are typically short in dataset publications (e.g., in Budzianowski et al., 2018; Wen et al., 2017) and design choices are not always argued for, our in-depth description can help researchers collect dialogue data in a simple way. 2) In our data analysis, we present novel operationalizations of quality metrics and show that small efforts in the data collection setup can lead to less bias in the dialogue data, thus making it more ecologically valid. Ecological validity is a notion introduced by de Vries et al. (2020) into the NLP community, which specifies "the degree to which [data] generalize[s] to naturally occurring scenarios" (de Vries et al., 2020). A dataset is ecologically valid, and thus allows for such generalizations, if it consists of (simulations of) human-machine interactions, that is, data which "reflect[s] the intents and linguistic phenomena found in real-world applications" (de Vries et al., 2020). 3) We release ***Cro**wdsourced **W**izard of Oz **D**ialogue dataset based on **S**ituated **S**cenarios* (*CROWDSS*), a dataset labeled with dialogue acts (DAs) which, to the best of our knowledge, is the first German task-oriented dialogue dataset. It can be used for a variety of NLP tasks like DA classification or response selection.

The paper is structured as follows. First, we describe what makes dialogue an efficient and flexible form of interaction and why dialogue is needed for SAs to be able to help users accomplish complex tasks in a natural way. Next, we review previous approaches to dialogue data collection, and we describe our own method, arguing for how our design choices fit in with our three goals (good-quality data, low-resource approach, feasibility in languages with limited crowd-worker availability). We briefly explain our annotation efforts, and go on to analyze our data regarding data quality. For this purpose, we compare our data to a sample of MultiWOZ (Budzianowski et al., 2018). De Vries et al. (2020) used that very corpus to demonstrate the presence of scripting and priming in datasets for NLP. Crucially though, MultiWOZ makes for an ideal comparison as it was collected in the same way as CROWDSS, with one key difference which is the stimuli used to elicit the dialogues (template-based vs. situated scenarios). Assessing pre-*script*-edness regarding task-relevant entities (scripting), lexical overlap between scenario and dialogue (priming) as well as diversity between

dialogues elicited from the same scenario (scalability), we can show that small investments in the collection setup can greatly improve the dialogue quality.

Dialogue as the Mode for Solving Complex Tasks

Current SAs including popular services like Apple's Siri or Amazon's Alexa do a good job at single-turn commands and simple multi-turn interactions. However, SAs could also support humans in complex tasks like *finding a restaurant* and *booking a table* there, *comparing shopping items*, or *searching large databases* (Asri et al., 2017; de Vries et al., 2020). To the extent to which SAs are already capable of assisting humans in such complex tasks, they are typically implemented in a simplified way: a *restaurant booking task* would, for example, be implemented as *slot filling*, which limits the exchange to the assistant *requesting information from the user* and the user *informing the assistant* (e.g., in Rasa—Bocklisch et al., 2017). In human-human interaction, on the other hand, such tasks would be collaboratively solved in a sophisticated way using *dialogue* (Stalnaker, 1978; Clark and Wilkes-Gibbs, 1986; Garrod and Anderson, 1987; Clark, 1996; Pickering and Garrod, 2004; Xu and Reitter, 2018).

Human dialogue is a flexible mode of interaction without a turn limit, enabling information to be exchanged dynamically, depending on the dialogue flow, a sudden change of mind, new incoming information, etc. The linguistic features that constitute dialogue make it a very efficient way of negotiating information while building common ground (Stalnaker, 1978; Clark and Wilkes-Gibbs, 1986; Clark, 1996; Xu and Reitter, 2018): already-introduced entities can be referred to with shorter expressions (anaphora, Poesio and Rieser, 2011); understanding from context makes it possible to omit superfluous elements (context sensitivity, ellipsis, Levelt and Kelter, 1982); mutual understanding is continuously displayed and monitored (back-channeling, Schegloff, 1982) and if need be enforced (error management), and there is an elaborate system for floor management (turn-taking, Oreström, 1983). Its flexibility and efficiency may explain why dialogue is so easy to process for humans (Fox and Jean, 1999; Branigan et al., 2011) and why it presumably is also their preferred format of SA interaction (de Vries et al., 2020).

In sum, the flexibility and efficiency of dialogue as a mode of interaction as well as a (presumable) general human preference for this type of communication indicate that SAs need to be able to handle natural dialogue if they are to assist humans in more complex everyday tasks. Dialogue as an interaction mode is, of course, not limited to task-oriented settings where SAs help users get a task done, but also social SA interactions can (and should) adopt a dialogical style. However, the following discussion of previous work solely focuses on data collection for task-oriented systems.

Previous Work

Dialogue datasets have been collected or generated in three different ways, distinguished by who is interacting with whom in the data collection/generation process.

Dialogue Data Collection Settings

In a **human-human** setting, two or more human users talk or write to each other, thereby generating a dialogical interaction. Since all interlocutors are human and thus dialogue-competent, the aforementioned dialogue-specific phenomena are ideally present in and can be learnt from data collected in this way. Yet, humans talk differently to machines than with their fellow human beings (de Vries et al., 2020, but see discussion below). Thus, simply employing existing human-human dialogue corpora (e.g., recorded dialogues from customer service interactions) may not be representative enough of human-machine interaction to improve current SAs. The human-human setting was applied, among others, for the MultiWOZ corpus (Budzianowski et al., 2018), as well as by Wen et al. (2017), Eric et al. (2017), Asri et al. (2017), Byrne et al. (2019) and for training Google Duplex (Leviathan and Matias 2018; Chen and Metz 2019), a SA which can interact with businesses on behalf of customers in the restaurant booking domain.

Apart from that, data has also been collected in a **human-machine** setting where a human participant interacts with a machine, for example in the setup of the second and third DSTC challenge (Henderson et al., 2013). While this type of setup may serve to improve an already-existing system, the improvement can only happen within the capabilities of that system (Wen et al., 2017; Budzianowski et al., 2018). Crucially though, natural dialogue-specific phenomena will only be present to a limited extent in data collected in this way. They will not be present in the machine turns at all. The users, then, may speak to the machine in a dialogical way, but this typically results in the machine's failure to react "dialogically", and, in the long run, in the users' downscaling their linguistic behavior to a level adequate to their machine interlocutor (see below).

Furthermore, to overcome data scarcity, data has also been generated by having **two machines** interact with each other, for example by Shah et al. (2018) and by Rastogi et al. (2020). The machine-produced interactions are typically an exchange of intents or DAs coupled with entities which are subsequently translated into natural language, for example by crowd-workers. While researchers are saved from the time-consuming and error-prone task of annotating the dialogues, data collected like that can in no way serve to improve a SA's dialogue competency as neither the "user" nor the system are dialogue-competent. Nonetheless, even datasets collected in this way are called "dialogue datasets" and are used to train dialogue systems.

Following this, if the goal is to enable machines to handle dialogue in a natural way, only data collected from human-human interactions seems adequate (Budzianowski et al., 2018; Byrne et al., 2019). Dialogue data from human participants has either been collected in an overt setup where both participants know they are interacting with another human being (Asri et al., 2017) or in the style of Wizard-of-Oz (Kelley, 1983).

Wizard-of-Oz Framework

The Wizard-of-Oz (WOz) framework aims at striking a balance between the naturalness of human-human interaction while still accounting for the observation that humans talk differently to

machines than with their fellow human beings. This is generally attributed to the fact that humans always adapt their language towards the recipient and what they perceive as the recipient's capabilities (Shatz and Gelman 1973; de Vries et al., 2020, but see discussion below).

Therefore, in a WOz collection, the participant playing the role of the user (here: *she*) is led to believe she is interacting with a machine. The machine, on the other hand, is enacted by a different participant, who plays the role of the assistant (Kelley, 1983). Naturally, the assistant (here: *he*) exhibits better natural language understanding (NLU) than machines and he produces a greater variety of answers to user requests than existing natural language generation (NLG) modules do (Byrne et al., 2019). In doing so, he generates system turns that are more representative of how human interlocutors would behave in similar dialogical scenarios (Merdivan et al., 2020). At the same time, he formulates his replies knowing that he is mimicking a machine, respecting crucial boundaries: he only "operates" within the task that he is "designed for", producing mostly task-oriented utterances and engaging less in social chat. On the other side, "[g]iven the human-level natural language understanding, [the participant playing the user] quickly realize[s she] can comfortably and naturally express [her] intent rather than having to modify behaviors as is normally the case with a fully automated assistant" (Byrne et al., 2019). Doing that, she generates user turns that are also more in line with natural dialogue as it would unfold between human interlocutors, employing, for example, context-dependent anaphora or elliptic structures. Yet, she also operates within the boundaries of an interaction with a machine which was designed for a specific task, and therefore avoids conversing all too freely with it (Byrne et al., 2019) as well as typical idiosyncrasies of human dialogue like back-channeling. It may, of course, be the case that an individual user participant becomes suspicious of the illusion of interacting with a machine, but this is not necessarily detrimental for the data, as this participant will still conform to the context of addressing a machine for accomplishing their task.

For human-machine interactions, there is a discussion regarding whether, in addition to the recipient-oriented talk, humans conceptualize machines as different, non-human entities, and (also) for that reason talk differently to them [de Visser et al. (2016) call this the "unique-agent hypothesis"]. Crucially, this would mean that human-machine interactions will always be of a different kind than human-human interactions. In contrast to this, the Media Equation theory, which originated from the *Computers as social actors* (CASA) framework (Nass et al., 1994; Reeves and Nass, 1996; Nass and Lee, 2001), argues that computers, too, are social actors, and that humans apply the same social rules and norms when interacting with them as they do when interacting with fellow human beings (de Visser et al., 2016). This theory can explain why people currently behave differently towards machines (i.e., due to their limited capabilities), but contrary to the "unique-agent hypothesis", it predicts that these differences will disappear with machines becoming more human-like. If machines take on different "personalities", possibly even user-adapted ones, users would still adapt their speech towards their machine interlocutors, but more so to the specific machine they are interacting with, to its "personal" style of behavior, rather than in a generic machine-directed way. For a discussion on how humanlike-ness is conceptualized and how it can potentially be reached by machines, refer to de Visser et al. (2016).

Since machines are a long way from being able to converse in a human-like way, whether or not human-machine interaction will always be fundamentally different from human-human interaction is beyond the scope of this paper. For now, we must reconcile the fact that current human-machine interaction lags behind human-human interaction (the latter being the model to strive for and learn from due to its advantages: flexibility, efficiency, suitability for complex tasks) and the knowledge that humans (at least for now) generally talk to machines differently. WOz provides a suitable solution for both these issues (de Vries et al., 2020) and has accordingly been applied in some of the major recent data collection efforts (e.g., Eric et al., 2017; Budzianowski et al., 2018; Byrne et al., 2019).

Additionally, WOz approaches try to simulate a context analogous to the eventual deployment context, which is "of utmost importance" for machine learning (de Vries et al., 2020). The Spoken Dialogue Systems community has always taken this simulation of the deployment context very seriously, actually striving for ecological validity and focusing on smaller, good-quality datasets (Rieser 2008; Rieser and Lemon 2011; Schlangen 2019). Lately, however, the focus has shifted to large-scale collections, both due to demands of huge data-driven models and thanks to the availability of online data collection platforms (e.g., Wang et al., 2012; Eric et al., 2017; Wen et al., 2017; Budzianowski et al., 2018; Wei et al., 2020). Such collections typically use templates to elicit data, which has been argued to induce researcher bias into the collected data, lowering its ecological validity (see below; de Vries et al., 2020). An interesting open question is then how we can optimally set up a WOz data collection on the web for a tradeoff between good-quality data and a low-resource collection setup.

(A)synchronous Interaction

Dialogue data from human interlocutors has either been collected from live interactions (Garcia et al., 2020) or in an asynchronous fashion (Eric et al., 2017; Wen et al., 2017; Budzianowski et al., 2018). Live interactions can be done with hired workers or with crowd-workers (e.g., over Facebook's ParlAI). In any case, a large pool of participants is required, because two participants have to be available at the same time to be paired up (Garcia et al., 2020), and in the case of crowd-workers there is a considerable risk of dropouts.

In asynchronous data collection, dialogues are collected on a turn-by-turn basis. This means that a participant's task only consists in continuing an ongoing dialogue with *one* turn. The dialogue including the new turn is then handed to the next available crowd-worker (in the opposite role) which can happen at any later point in time. This simplifies the setup and eliminates the need that two crowd-workers be available at the same time. Despite the involvement of multiple rather than two interlocutors, turn-based data collection has been shown to

generate coherent data (Wen et al., 2017; Budzianowski et al., 2018). All-important coherence in dialogues is ensured by having participants read the dialogue history before they respond to it, as well as through a unifying scenario for the user participants and in some cases also a unifying user persona (Jonell et al., 2019).

Modality

Dialogue data collections further differ regarding whether the data is collected in spoken or written modality. As language is not only recipient-dependent, but also modality-dependent (Serban et al., 2018), data from spoken interactions is, naturally, most representative for the smart *speech* assistant domain. Collecting spoken rather than written data requires a more complex data collection setup in which participants record their utterances with microphones. These utterances, then, need to be transcribed by an automatic speech recognition (ASR) module, which can be error-prone (Asri et al., 2017).

While the spoken modality makes sense for live interactions, for asynchronous data collection, it poses the non-trivial problem of how the dialogue history (and even the scenario) should be presented to new crowd-workers who are continuing the dialogue. One option would be to present the dialogue history either completely in written form or with the last turn being synthesized. Yet, this would result in a mix of modalities where some or all of the dialogue history is in text form, but the crowd-workers are asked to phrase the dialogue continuation using their voice. A second option would be to synthesize the whole dialogue history using two different voices, but that would be rather strange for the crowd-workers as one of the two voices is supposed to be their own. Furthermore, only listening to the dialogue may make it more difficult for crowd-workers to familiarize themselves with the dialogue that they should continue, since with text one can more easily revisit what has already been discussed. Thus, for asynchronous data collection, using the spoken modality not only means more technical overhead, but also a modality mix which likely does not result in higher data representativity for the SA domain than completely replacing the spoken modality with the written one. Therefore, asynchronous datasets have been collected in the written modality (Eric et al., 2017; Wen et al., 2017; Byrne et al., 2019 [note that they also collect spoken data for the same dataset]; Budzianowski et al., 2018), often without even discussing modality as a factor.

Furthermore, with online chats having become a popular means of interaction, the line between the written and spoken modality seems to have become blurrier (Nishimaki 2014). Chats exhibit some of the features that previously had been proprietary to spoken language (Dürscheid and Brommer, 2009; Nishimaki, 2014), including, but not limited to shorter, spontaneously produced units, ellipsis, colloquial style, (near) synchronicity, and dynamic error management. At the same time, chat platforms like WhatsApp or Threema (where the modality focus traditionally was on written chats) also provide the possibility to send voice messages, and users often mix the written and spoken modality. Interestingly, when switching to the spoken modality, the messages are decidedly non-dialogical not permitting any form of interaction or feedback until the message is recorded, sent and listened to. In any case, collecting data in an asynchronous way (i.e., typing a message into a response field after viewing the dialogue history which is typically presented in differently-colored text boxes, aligned on either side of the window), could remind crowd-workers of chat interactions and stimulate a more oral-style language.

Summary

The main distinction regarding dialogue data collection approaches comes down to who is interacting with whom in the collection process, and, along with it, which purpose the dataset should serve: train better, dialogue-competent NLP models (human-human setting), improve an existing system (human-machine setting), or overcome data scarcity (machine-machine setting). Most datasets collected from human-human interactions apply the WOz framework in order to obtain data that is not only representative of natural dialogue, but also of a human-machine interaction context. WOz data collections can be distinguished regarding whether the participants interact live or asynchronously, and along with it, whether they do so in the written (both) or spoken modality (only live interaction).

Our Approach

Based on the preceding overview, we collect written, task-oriented dialogues from human-human interactions in the WOz framework and an asynchronous fashion, aiming at a tradeoff between good-quality data and a low-resource collection setup. The dialogues are collected with the help of crowd-workers on AMT. We propose to collect data asynchronously as it is in accordance with our low-resource goal. Crucially, this design choice is also suitable for collecting data in languages that are less represented on AMT than English, as pairing up two simultaneously-available crowd-workers becomes obsolete. The asynchronous approach entails that we collect written data, which may lower the data's representativity. However, we argue that we can mitigate this issue by designing materials which stimulate an oral context (see below). Furthermore, we collect data in batches (i.e., all first turns of all dialogues in the first batch, then, all second turns of all dialogues in the second batch, etc.; see below) which further simplifies the collection setup.

We follow a similar approach to the one used to collect MultiWOZ (Budzianowski et al., 2018), as we crowdsource written dialogues from human interlocutors in an asynchronous WOz setup. However, de Vries et al. (2020) used MultiWOZ to exemplify the presence of scripting and priming in NLP datasets, arguing that this lowers its ecological validity. Scripting and priming in the data can only be induced by the materials used to elicit the dialogues. Therefore, rather than using template-based scenarios like in MultiWOZ, we propose to use *situated scenarios* (see below), aimed at collecting good-quality data despite the low-resource setup.

MATERIALS AND EQUIPMENT

In order to collect task-based dialogues, there needs to be a task that the crowd-workers playing the user need assistance with, and

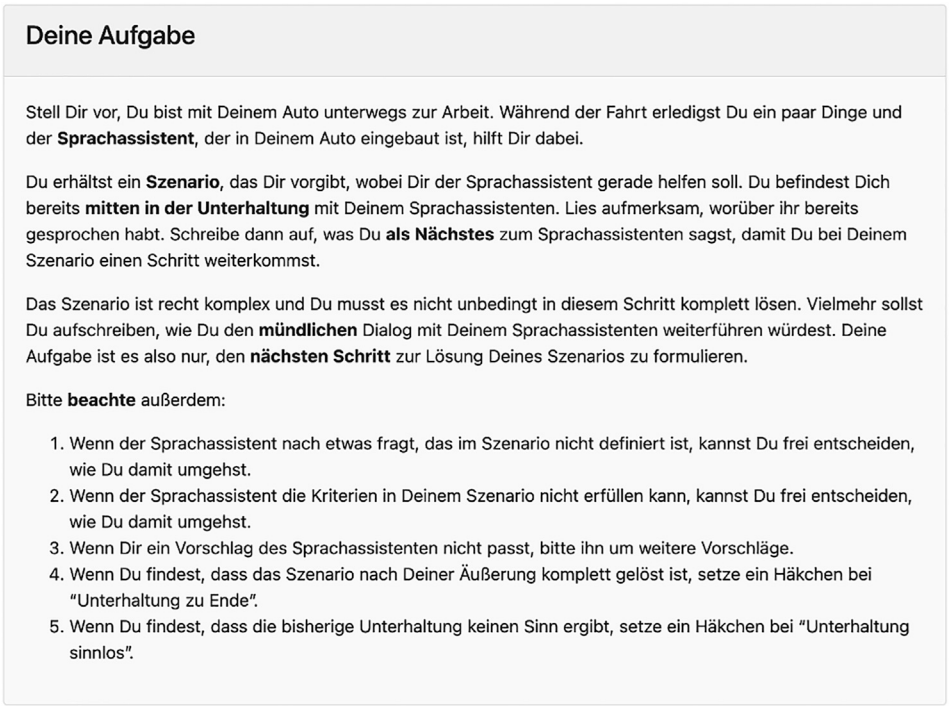

FIGURE 1 | Instructions familiarizing the user participants with their task.

that they should solve collaboratively with the crowd-workers playing the assistant. The user participants (UPs) got instructions for their role and a scenario which specified the task for a given dialogue. The assistant participants (APs) also got instructions for their role and a database which they could query to help the UPs accomplish their task.

Instructions for User Participants

The instructions for UPs introduced a simulated situation: they were driving with their car to their workplace (see **Figure 1**)[1] and while doing so they should solve a specific task with the help of an intelligent in-car SA. That task was detailed in the scenario (see below). It was stressed multiple times that the UPs should carry out *only one step* in order to come closer to solving the task (see bold words in **Figure 1**). This should prevent UPs from taking too lengthy turns, which would not be representative of spoken dialogue, where pieces of information are typically negotiated step by step. Also, the instructions put emphasis on the fact that it was a simulation of an *oral interaction* taking place in a hands-free environment. The UPs were asked to read what they had already *talked* about with their SA and to write what they would *say* next to the SA if it were an *oral* dialogue.

The instructions further detailed a few other points that the UPs should consider, including that they could freely choose how to react if the SA could not meet their request or if they were presented with a question that they could not answer by relying on the scenario. They were furthermore encouraged to evaluate offers from the SA according to their own liking. These points were included to give the UPs more freedom, which should lead to more diverse dialogues. Furthermore, the UPs were instructed to tick a box, 1) if they deemed that the dialogue was completed, and/or 2) if they thought that the preceding dialogue history was incoherent. This built-in, crowd-sourced dialogue validation mechanism could make scaling easier. Lastly, they had the option to leave us a comment.

Situated Scenarios for User Participants

The task that the UPs should pursue in their dialogue (*finding a restaurant* and, in some scenarios, also *booking a table*, see below) was specified in a scenario. Asynchronous data collection has been shown to generate coherent dialogues (Wen et al., 2017; Budzianowski et al., 2018) as long as all UPs working on the same dialogue are presented with the same scenario and the dialogue history up to their turn. Template-based scenarios can help achieve coherence, but they arguably offer the participants too fixed of a script to solve the task and prime their lexical choices (de Vries et al., 2020). As our first goal was to collect good-quality data that is not only coherent, but also diverse and natural, our scenarios had to specify the right amount of information to, on the one hand, achieve a unified intent (in the interest of coherence), but, on the other hand, give enough freedom to the UPs, making sure that the dialogues are not completely pre-*script*-ed (in the interest of diversity and naturalness).

[1]Find translations of all German-language figures, tables and examples in the **Supplementary Material**.

In order to achieve all of that, we designed *situated scenarios* to tap into the participant's situated knowledge of *restaurant booking*. Situations such as *finding a restaurant* or *booking a table* are represented in our brain as complex simulations of perceived situations (Barsalou, 2009), which include information regarding relevant people and objects, typical actions, background settings as well as introspections, intentions and emotions. We can tap into this situated knowledge without having to spelling out every single detail of a situation (as do template-based scenarios), but using minimal lexical or visual cues (McRae et al., 2018). For example, it can be assumed to be common knowledge that some restaurants do not allow pets and hence "your dachshund Benno should also be allowed to tag along" (see Example 1 below) should be enough of a cue to let UPs know they should consider this criterion in finding a restaurant. Situated knowledge also includes what typically motivates a situation: people do not just book a restaurant matching a set of criteria for the sake of booking a restaurant, but they engage in such a situation with a specific motivation in mind. Therefore, to make the scenarios more realistic, we also included some background information about the person that the UPs were simulating as well as their motivation (e.g., in Example 1, the holidays in Brittany and the there-acquired love for French cuisine as background; taking out the girlfriend as motivation).

Example 1 shows an example scenario including a booking task:

> "Example 1: Die letzten paar Urlaube hast Du mit Deiner Freundin in der Bretagne verbracht. Da hast Du die Küche dieses Landes lieben gelernt: Baguette, Croissants, Käse und Rotwein... Heute Abend möchtest Du mit ihr essen gehen. Euer Dackel Benno sollte auch mitkommen dürfen. Der Abend sollte nicht zu teuer werden, es muss aber auch nicht das billigste Restaurant sein. Finde ein passendes Restaurant und buche einen Tisch für Euch. Bringe außerdem die Adresse des Restaurants in Erfahrung."
>
> English translation: "*You've spent your last couple of holidays with your girlfriend in Brittany. You've got to love the cuisine of that country: baguette, croissants, cheese and red wine… Tonight you want to take her out to dinner. Your dachshund Benno should be allowed to tag along. The evening shouldn't be too expensive, but it need not be the cheapest place either. Find a matching restaurant and book a table for you guys. Also, inquire about the restaurant's address.*"

Further, the information that we chose to provide for finding a restaurant should not jeopardize the naturalness of the dialogues on the level of wording. The scenario in Example 1 nicely demonstrates our efforts to avoid lexical priming in the dialogues. It specifies three criteria for finding a suitable restaurant (*French cuisine, dog-friendly, medium price range*) and two for booking a table there (*tonight, two people*). Crucially, none of these criteria are *explicitly* given *in full detail*: *French cuisine* is hinted at by the vacations in Brittany as well as typical French dishes, *pet-friendly* is paraphrased by means of "your dachshund Benno should also be allowed to tag along" (note that, in German, the word *Hund* [dog] is not used), *medium price range* can logically be deduced from "not too expensive, but not too cheap either", *tonight* is explicitly specified, the exact time, however, is not, and, lastly, *two people* must be derived from "you and your girlfriend". We used vague language and circumlocutions for all the criteria in the scenarios, in order to afford crowd-workers the opportunity to phrase their utterances in their own words, rather than priming them into using the exact same words as in the scenarios.

In total, we used ten different handcrafted scenarios as seeds for the dialogues (see Batch-wise data collection). Five of them only concerned *finding a restaurant*, and the other five also required *booking a table* there. Each scenario contained three search criteria, which mirrored the search criteria in the database that the APs had at their disposal (see below). Thus, the scenarios were so detailed that the UPs were likely to go about solving them in multiple steps, that is, engaging in a dialogue.

In sum, our situated scenarios describe the task in a way that should reduce priming while still cueing the target situation, and at the same time permit some degree of freedom for the UPs, and we expect them to yield coherent, yet diverse, and natural dialogue data.

Instructions for Assistant Participants

For the APs, the instructions explained that their task was to play the role of an in-car SA and help a human driver find and book a restaurant (and *only* do that; see **Figure 2**). They were instructed to carefully read the dialogue history, extract relevant information from it and use it to query a simple database which we provided (see below), to find a restaurant matching the user's request. The APs also had the option to mark dialogues as completed or as incoherent, as well as leave us a comment.

APs were further informed about their capabilities as SAs: They could simulate making bookings, calling venues as well as navigating there. To simulate bookings, APs continuing the dialogues past the fourth turn (see Batch-wise data collection) saw three additional text fields in their graphical user interface (GUI; see below), where they were asked to enter any booking information (*name of restaurant*, *day* and *time of booking*, *number of guests*, respectively) as soon as the users had settled on one of them. Bookings were always successful, as they only consisted in filling in these fields. The database was the APs' only source of information. Thus, if a user asked for reviews about a restaurant, this request should be considered as "out of scope", as reviews were not part of the database (see below). Again, we put emphasis on the fact that the APs were engaging in what should be an oral interaction.

Database for Assistant Participants

The database (see **Figure 3**) that the APs had at their disposal was implemented as a simple combination of HTML forms and client-side JavaScript which was integrated into the HTML code for the GUI (see below). It consisted of 200 different restaurants, where each had a unique name and was defined regarding the search criteria cuisine (*American, Chinese, French, German, Italian, Mexican, Turkish, vegan*), location (*downtown, North, South, East, West, countryside*), price range (*cheap, moderate, expensive*) as well as an additional feature (*live music, wheelchair accessible, dog-friendly, featuring garden/terrace, accepts credit cards*; note that, for simplicity, these features were mutually exclusive). Furthermore, each

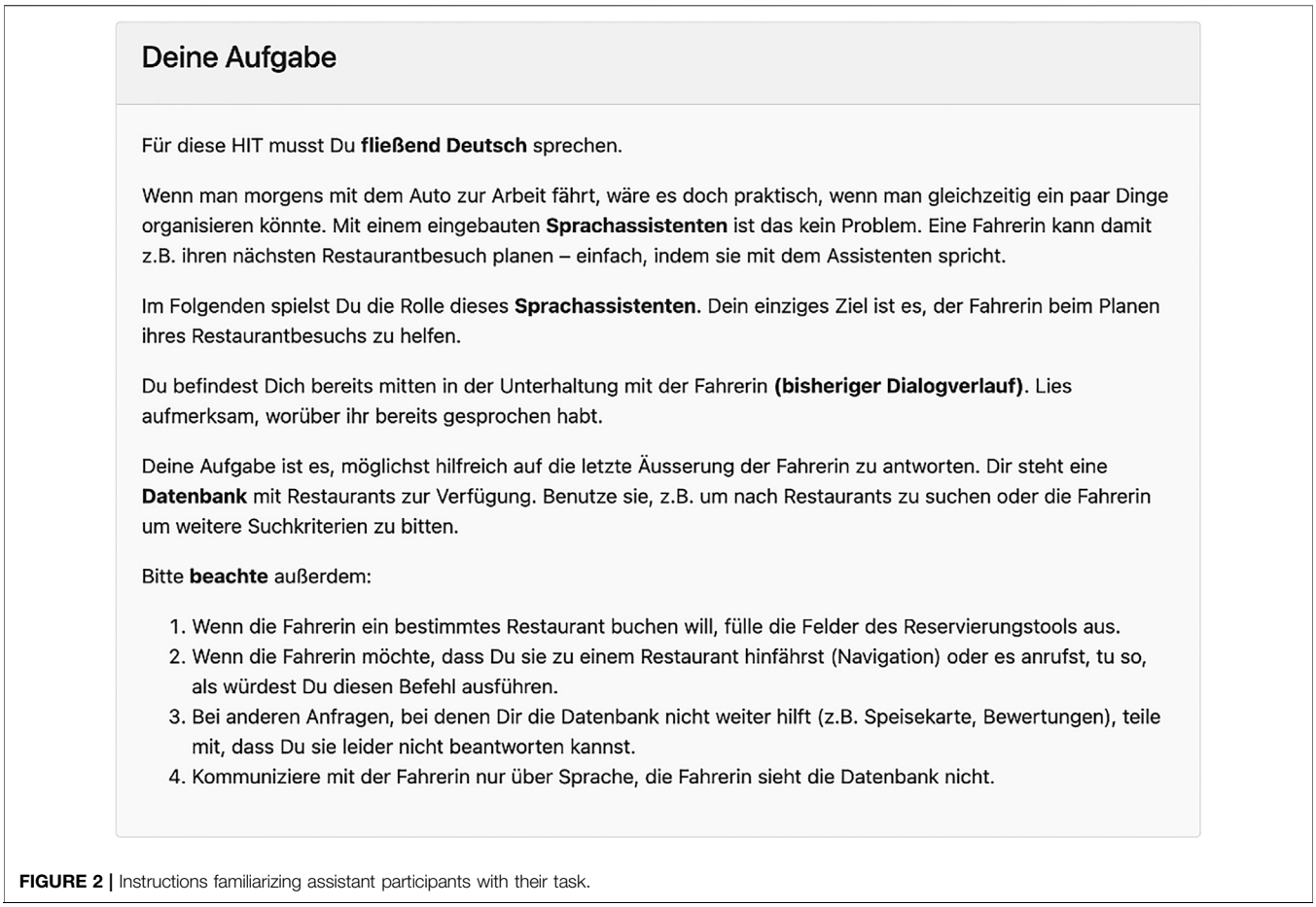

FIGURE 2 | Instructions familiarizing assistant participants with their task.

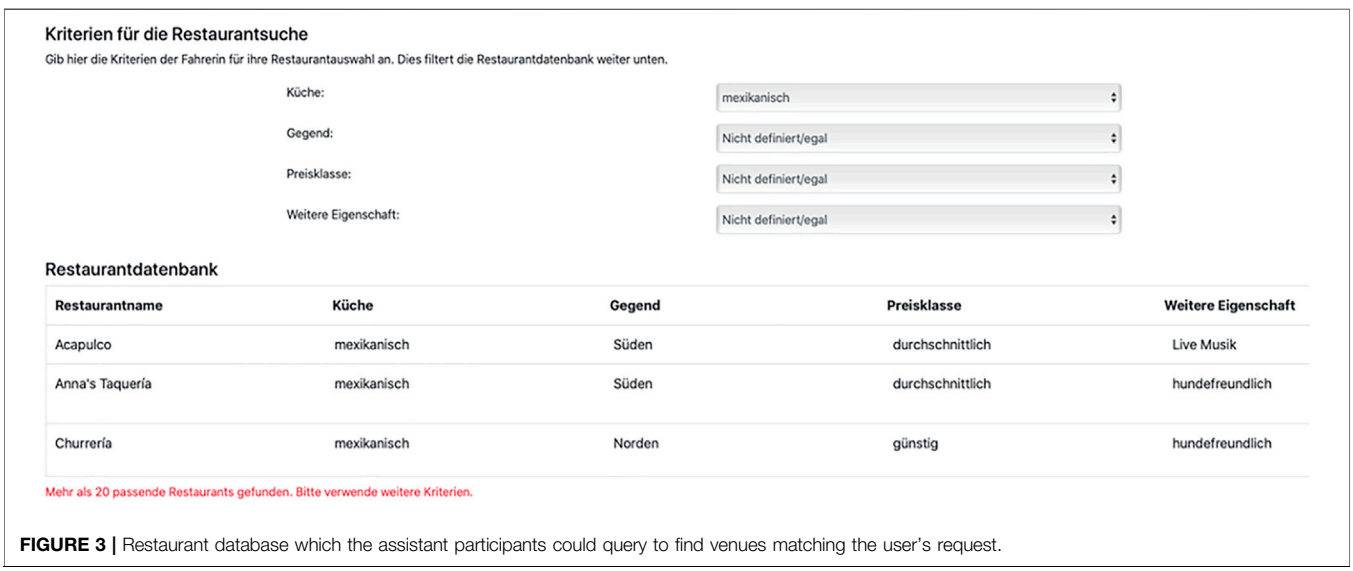

FIGURE 3 | Restaurant database which the assistant participants could query to find venues matching the user's request.

restaurant featured an address including an indication of the distance from the users' current location which the UPs could request from the APs (see Example 1). The database could be filtered using any combination of the search criteria.

Presupposing that UPs would extract the intended search criteria from the scenario, request them in their utterance, and that the APs would use them to filter the database, four out of the ten scenarios were designed to lead to multiple entries in the

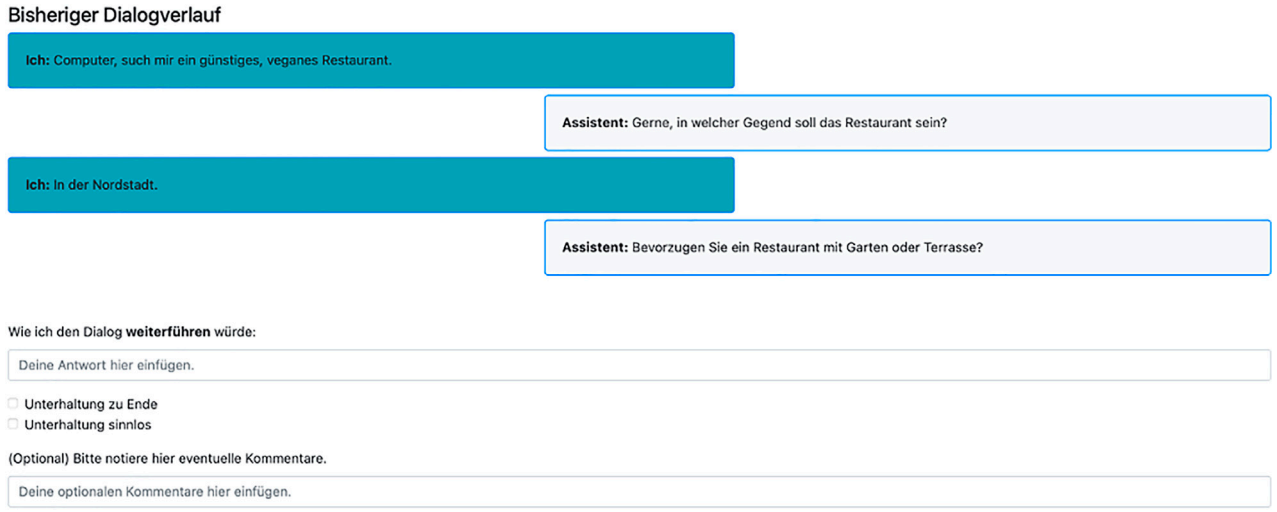

FIGURE 4 | Graphical User Interface for user participants with dialogue history, a response and optional comment field as well as end/spam check boxes.

FIGURE 5 | Map of fictional city "Bornberg" which was embedded into the Graphical User Interface of both participant groups, establishing common ground between them.

database, another four would instead lead to one entry each, and the last two to zero entries. This should again produce more diverse dialogues.

As the dialogues were gradually continued (see Batch-wise data collection), APs always saw the database output with the most recently-applied filters from the previous AP turn, but they could always modify them to yield different search results.

The architecture of AMT is centered around the publication of batches of *human intelligence tasks* (HITs). One batch contains one or more HITs, each of which can be assigned to one or multiple crowd-workers. The platform offers templates for common crowdsourcing setups (e.g., sentiment analysis) for task completion on-site. Alternatively, it can be used to only recruit participants for task completion on an external website. Further, it features an HTML editor where researchers can design their custom GUI for task completion on-site. In pursuit of our low-resource goal, we opted for this possibility and designed two complete GUIs inside AMT's HTML editor:

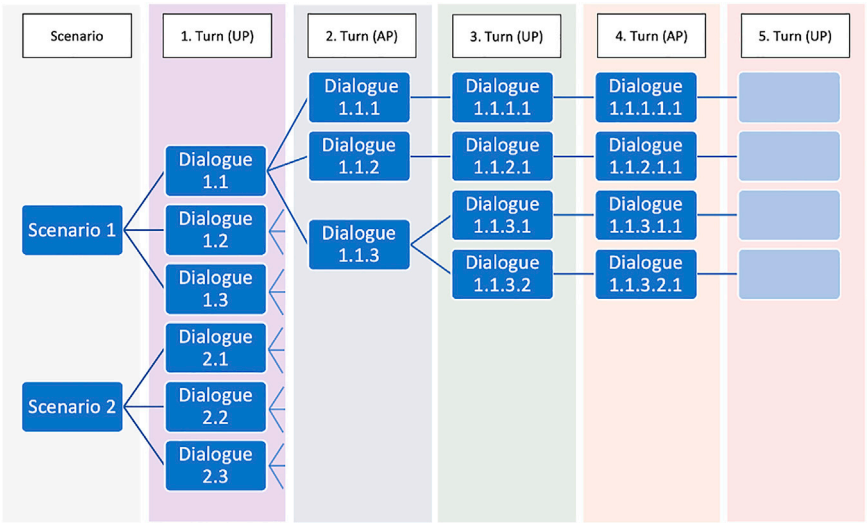

FIGURE 6 | One-to-many ratio showing how one dialogue seed was used to elicit multiple dialogue continuations.

one for UPs (see **Figure 4**), and one for APs including the searchable database.

For common grounding, participants of both roles further saw a map of the fictional city "Bornberg" in their GUI (see **Figure 5**). The map featured a pin indicating the user's current location. This should enhance the situatedness of the task of *finding a restaurant* in a specific geographic environment, by providing an additional cue in the visual modality.

METHODS

Having prepared instructions, scenarios, the database and the GUIs, we launched our data collection. We collected the data in batches, which is fully in line with AMT's architecture.

Batch-wise Data Collection

In batch 1, we presented the instructions and a scenario to UPs, who (according to their role) were assigned a scenario to pursue, and who had to initiate the dialogue. Applying a one-to-three ratio, we assigned each of the ten scenarios to three unique workers, to collect first utterances for each scenario, thus generating 30 dialogue beginnings in batch 1 (see **Figure 6**)[2]. Having collected them, we ran a post-processing script on the data, which spell-checked the utterances (relying on spaCy and pyspellchecker; Honnibal et al., 2020; Barrus 2019), corrected common misspellings (like wrong lower-casing of the polite form *Sie*) and flagged duplicate, empty or overly long answers (twelve words or more) for manual review. Our asynchronous batch-wise collection setup allowed us to do this offline and under no time pressure.

Next, in batch 2, we used the 30 dialogue beginnings as seeds to collect dialogue continuations (i.e., the second turn) from APs. The APs were presented with instructions, the database and a dialogue history which, at this point, only consisted of one previous turn. We again applied a one-to-three ratio and assigned each beginning to three unique APs, thus collecting 90 two-turn dialogues by the end of batch 2. We excluded 7 of them, as they had been marked as incoherent, and ran our post-processing script on the remaining 83 dialogues, manually reviewing flagged ones.

In batch 3 we then collected dialogue continuations from UPs, who were now instructed to familiarize themselves with the dialogue history as well and to continue the dialogues. Further, if they were working on a scenario which required *booking a table*, they were only now presented with this additional task as we wanted to prevent information overload in batch 1. For batch 3, we did not follow the one-to-three ratio between dialogue seeds and new continuations, but we assigned about a third of the 83 dialogues (30) to two UPs each (one-to-two ratio) and the rest (53) to only one UP each (one-to-one ratio). This was an arbitrary choice to limit the overall size of the dataset, as we primarily wanted to gain experience in collecting dialogue data in an ecologically-valid way, pathing the way for future domain-specific collections. At the end of batch 3 we obtained 113 unique dialogues. For batch 4 and onward, we continued assigning these 113 dialogues to participants following a one-to-one ratio, until they were complete. A dialogue was complete when one of the participants marked it as completed. Such dialogues were reviewed by us and, if complete, excluded from subsequent batches. Post-processing was performed after each batch.

Participants

In total, 57 unique participants contributed to the dialogues, of which 34 as UPs and 23 as APs. All crowd-workers had to be

[2]Note that the dialogue ids do not correspond to the dialogue ids in the dataset, as numbering in the dataset started with the first turn rather than with the scenario.

TABLE 1 | Comparison of CROWDSS with similar datasets (partially from Budzianowski et al., 2018; numbers for MultiWOZ and the dataset collected for Eric et al., 2017 are for the training split, for FRAMES the division is not specified, for CROWDSS the numbers refer to the whole dataset).

Metric	MultiWOZ Budzianowski et al. (2018)	FRAMES Asri et al. (2017)	Dataset collected for Eric et al. (2017)	CROWDSS
# Dialogues	8,438	1,369	2,425	113
Total # turns	113,556	19,986	Not provided	897
Total # tokens	1,490,615	251,867	Not provided	9,487
Avg. turns per dialogue	13.46	14.60	5.25	7.94
Avg. tokens per turn	13.13	12.60	9	10.1
Total unique tokens	23,689	12,043	1,601	906
# Slots	24	61	15	12
# Values	4,510	3,871	284	26

located in Germany. Additionally, they were asked to only participate if they were fluent in German. Some participants had to be excluded from further participation because of their non-fluent level in German, which became apparent when our post-processing script flagged their utterances due to, for example, misspellings. Participants should have an approval rate greater than 95% on AMT and we paid them above German minimum wage.

To uphold the WOz illusion, crowd-workers were consistently excluded from participating in the opposite role using AMT's built-in worker management, but not from participating in subsequent batches of the same role. However, in order to allow for more worker diversity in the data, we split batches into (up to eight) sub-batches and used "intra-role" worker exclusion, meaning that, for example, a UP from sub-batch 3.2 could not participate in sub-batch 3.3.

Annotation

We annotated our dialogues with DAs relying on the "Hierarchical Schema of Linked Dialog Acts" proposed by Pareti and Lando (2019). DA annotation schemes are typically developed either for human-human interaction (e.g., Bunt et al., 2010), or to meet domain-specific engineering requirements (e.g., the scheme used for MultiWOZ). Pareti and Lando's scheme bridges this gap as it was specifically developed to account for human-machine interaction, but in a broadly applicable, domain-agnostic way. It features "useful categories that can help [. . .] understand the human dialog input as well as generate a suitable machine reaction" (Pareti and Lando 2019). The schema is hierarchical with three levels of granularity (but can be extended, also with lower-level domain-specific tags if needed; see **Table 1**), where a full tag is composed of three sub-tags, one for each level, for example *request.instruct.task*. Further, the schema is expectation-based, meaning that the tags should be "informative of the [most salient] conversational expectations at each given point in the dialog" (Pareti and Lando 2019). Hence, for example, offers made by the assistant to the user fall into the high-level category of *requests*, as they create an expectation on the user to *accept* (or *reject*) it. In the original proposal, DAs are further linked with each other (beyond simple order), but for reasons of simplicity, we did not use this feature.

Conducting an iterative annotation pilot involving three trained linguists, we generally deemed it intuitive to find a tag from the schema for a given sequence in our data, and would typically also agree on the choice of the tag. However, as the schema does not provide descriptions of the individual tags, we created our own annotation guidelines. We share them (see **Supplementary Material**), as this may help other researchers annotate dialogues with the same scheme. We did not modify the scheme, but disposed of some tags that were not needed for our data (e.g., the original scheme contains further *assert* tags for *opinions* and *elaborations*). Annotation of all dialogues was then performed by two of the three annotators using Doccano (Nakayama et al., 2018; see **Figure 7**). The dialogues were pre-segmented into utterances, but we did not restrict the number of consecutive tags that the annotators could assign to an utterance. This resulted in some cases (19 out of 897 utterances) in a different number of tags assigned by the annotators to the same utterance. For example, one annotator would use one tag (e.g., *request.instruct.task*), whereas the other would further segment the utterance and use two (e.g., *social.greetings.opening, request.instruct.task*). In order to compute inter-annotator agreement (Cohen's kappa), we considered the smallest annotated unit (by either annotator) and, when needed, we doubled the tags assigned by the other annotator to ensure an equal number of tags assigned to each utterance. The resulting inter-annotator agreement (excluding the dialogues annotated during the pilot) was very high at 0.91. This is encouraging, as it suggests that crowd-workers could perform this task in large-scale collections, increasing the potential of scalability. Finally, the few inconsistencies were reconciled.

For the data quality analyses, we also annotated both dialogues and scenarios with entities, labelling all entities that are needed for *finding a restaurant* (`cuisine, location, price range`), *booking a table* there (`number of people, day of booking, time of booking`) or that could be requested by the user and retrieved from the database (`address`). Additional features (`live music, wheelchair accessible, dog-friendly, featuring garden/terrace, accepts credit card`) were annotated as Boolean entities. Annotation was performed in Doccano by one trained linguist. For each entity, we also saved the corresponding surface form in the scenario or dialogue and additionally normalized that surface form into entity categories, leaving us with, for example, the surface form

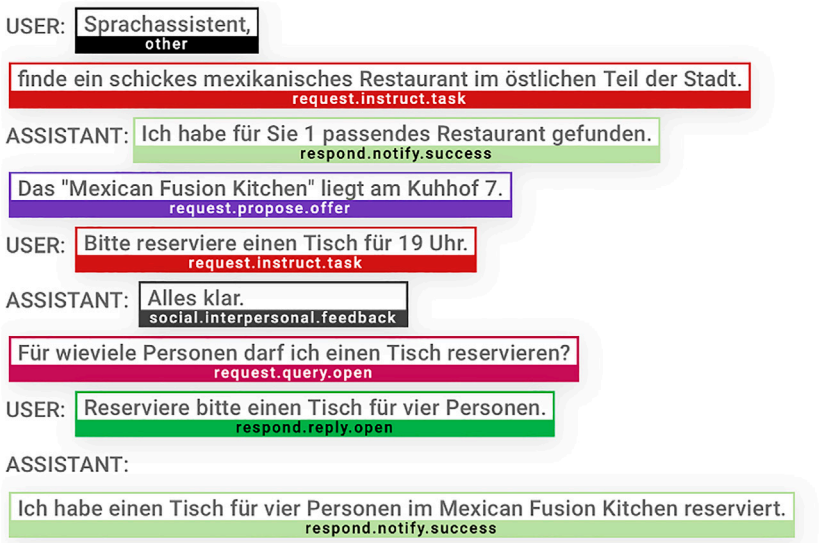

FIGURE 7 | An example dialogue annotated with dialogue acts in Doccano.

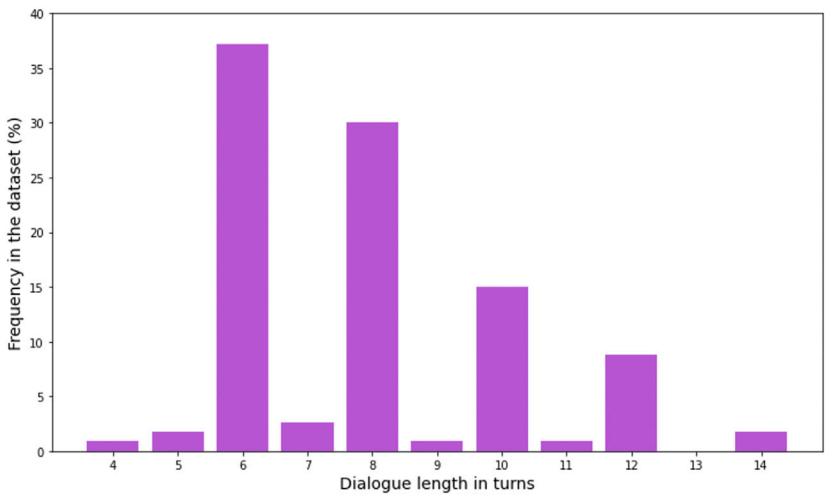

FIGURE 8 | Distribution of dialogue length in number of turns in CROWDSS.

"Küche des Reichs der Mitte" for the entity category *Chinese* for the entity `cuisine`.

Since our data quality analyses are partly comparative, we performed the same entity annotations for a sample of dialogues including the corresponding scenarios from MultiWOZ (see below).

RESULTS

In total, we collected 113 dialogues with a mean length of 7.94 turns per dialogue (*SD* = 2.13). The shortest dialogue consisted of four turns, the longest of 14 turns. Mean turn length was 10.1 tokens (*SD* = 3.98). The type-token-ratio (TTR) is 0.09 for the whole dataset and 0.71 on average for single dialogues. The two numbers differ substantially, as TTR is very sensitive to text length. In contrast, the Measure of Textual Lexical Diversity (McCarthy and Jarvis, 2010), a measure which avoids correlation with text length, is 67.72 for the whole dataset and on average 70.58 for single dialogues. **Table 1** presents a comparison between CROWDSS and similar datasets. **Figure 8** shows the distribution of dialogue length in number of turns in CROWDSS. The high number of dialogues ending with an assistant turn (even-number dialogue length) rather than with a user turn is due to the fact that APs typically marked dialogues as completed as soon as they had filled in all booking

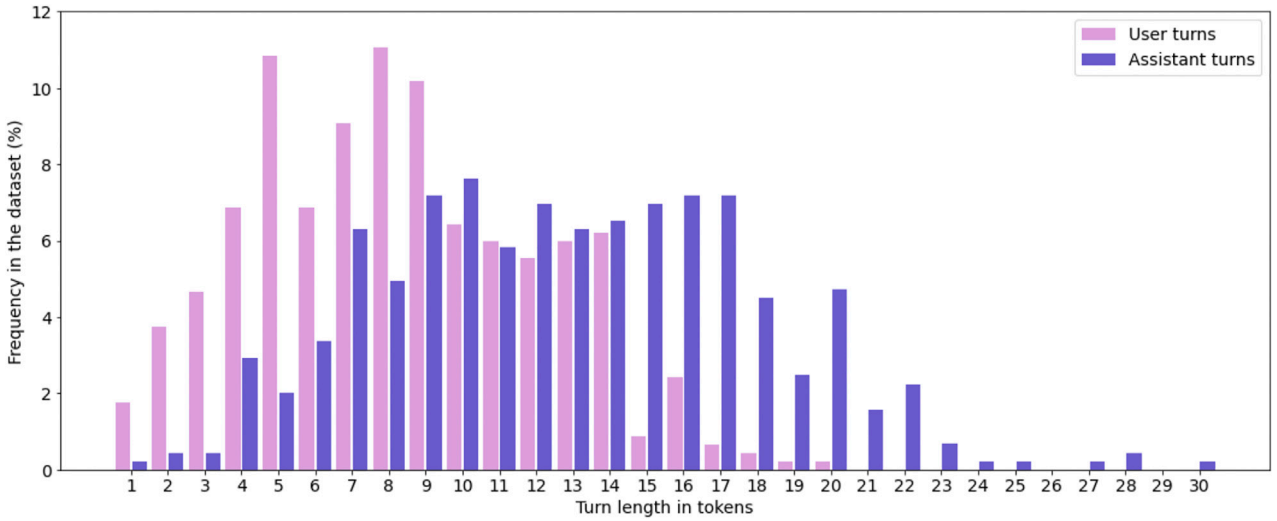

FIGURE 9 | Distribution of turn length in number of tokens for user and assistant turns in CROWDSS.

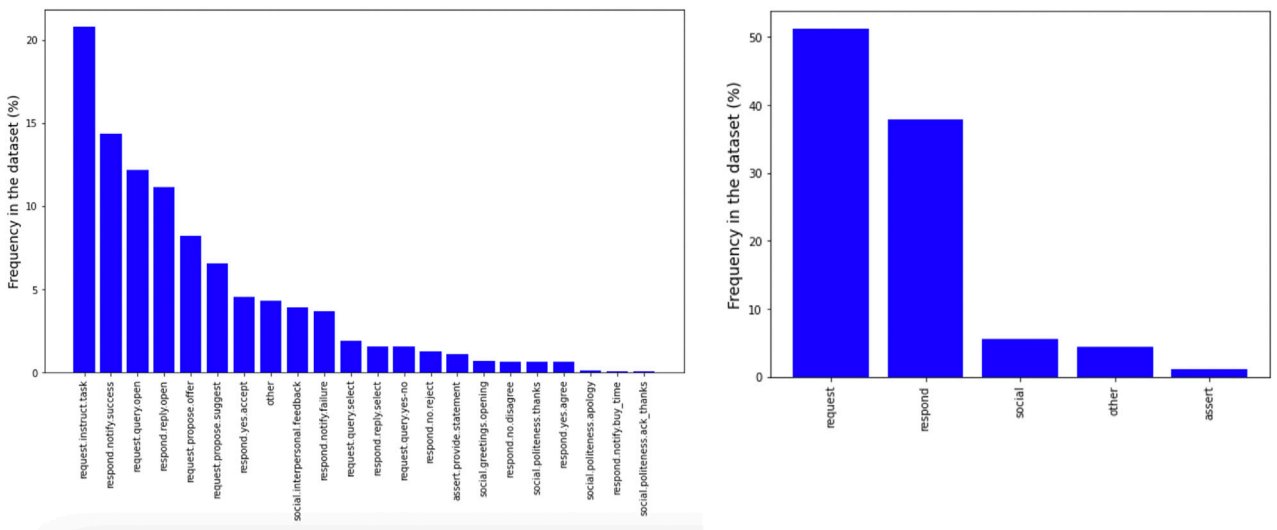

FIGURE 10 | Distribution of dialogue acts and high-level dialogue acts in CROWDSS.

fields and we manually confirmed that this was the case. **Figure 9** shows the distribution of turn length in tokens for user and assistant turns, respectively. **Figure 10** shows the distribution of all dialogue acts and of high-level dialogue acts, respectively (see Annotation). Finally, **Figure 11** shows the distribution of number of dialogue acts per turn.

Data Quality Analysis

Our first goal was to collect good-quality data that is both coherent, diverse and natural. In terms of coherence, we observe that our built-in spam detector was rarely used by the participants, suggesting that they had no problems making sense of the ongoing dialogue and continuing it with a meaningful next turn. Also, the optional comment field was not used to indicate any trouble regarding that. Furthermore, having annotated CROWDSS, it is our impression that the dialogues are coherent.

Diversity and naturalness, then, can be subsumed under the concept of ecological validity (see Introduction). A dataset can deviate from an ideal ecologically-valid methodology in five ways: 1) if it relies on synthetic language, or, given that the dataset contains authentic language, 2) if it was collected using an artificial task without real-world correspondence, 3) if it does not come

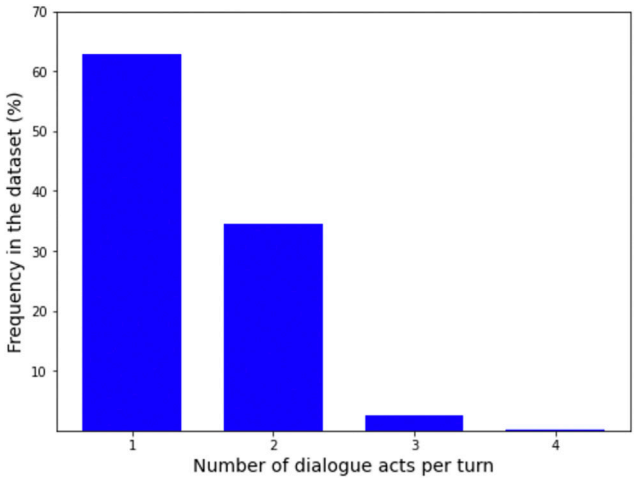

FIGURE 11 | Distribution of number of dialogue acts per turn in CROWDSS.

from users who may eventually benefit from a SA capable of a given, meaningful task, 4) if it contains single-turn interactions lacking the "conversational aspect", or 5) if template-based scenarios were used for data collection, which can lead to scripting and priming in the dataset (de Vries et al., 2020). WOz setups have typically aimed at sidestepping most of these issues. The use of template-based scenarios, however, can potentially be critical and affect the ecological validity by pre-*script*-ing the dialogues too much and by priming the participants' lexical choices.

We thus focus our data quality analysis on investigating the presence (or, ideally, the absence) of scripting and priming in CROWDSS. For that, we compare our dataset to the English-language MultiWOZ dataset (Budzianowski et al., 2018). We find that despite being in a different language, MultiWOZ makes for an ideal comparison. First, because de Vries et al. (2020) used that very dataset to show the presence of scripting and priming. Second, MultiWOZ was collected in the same way as CROWDSS (even the domain is the same) but with the potentially critical difference regarding the stimuli: template-based scenarios in MultiWOZ, handcrafted situated scenarios in CROWDSS. As for the language difference, we cannot exclude the possibility that the vocabularies for the task at hand are not equally rich in the two languages (e.g., in an extreme case, the English vocabulary could be so restricted that participants would make a given lexical choice, irrespective of a stimulus priming it or not). However, we argue that it is reasonable to assume similar vocabularies, given the closeness of English and German as well as the everyday nature of the task. For the analyses of lexical overlap (see Priming), we lemmatized all tokens so that different German case endings could not lead to an excessively low rate of overlap for German.

We first extracted a sample from MultiWOZ (n = 10,438 dialogues spanning multiple domains) to match the size of our dataset (n = 113). This was done by computing a random same-size sample from all single-domain dialogues in the restaurant domain. We observed that the restaurant dialogues in MultiWOZ

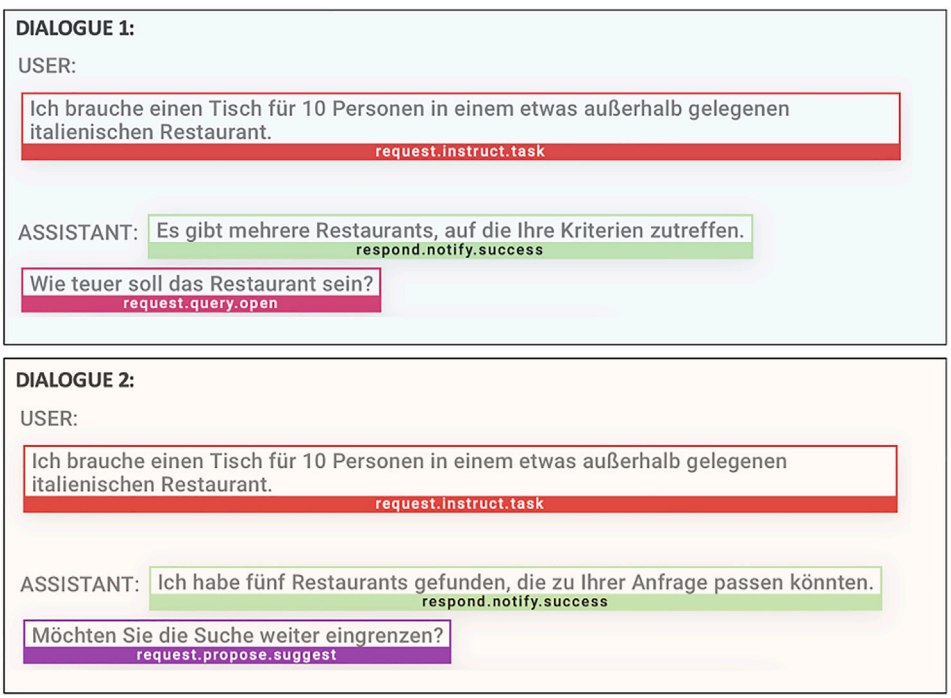

FIGURE 12 | Dialogue Act Sequence Similarity between turns elicited by same dialogue beginning: the two second turns share the first dialogue act and differ regarding the second one resulting in a sequence similarity of 0.5.

TABLE 2 | Entities in scenario and dialogue of MultiWOZ example dialogue.

Location	Order	Entity	Category	Surface form
Scenario	1	`location`	*south*	"south"
Scenario	2	`cuisine`	*international*	"international"
Scenario	3	`price_range`	*expensive*	"expensive"
Scenario	4	`cuisine`	*indian*	"indian"
Scenario	5	`number_of_people`	*4*	"4 people"
Scenario	6	`time_of_booking`	*T16:15*	"16:15"
Scenario	7	`day_of_booking`	*wednesday*	"wednesday"
Scenario	8	`time_of_booking`	*T15:15*	"15:15"
Scenario	9	`reference_number_req`	*reference_number_req*	"reference number"
Dialogue	1	`location`	*south*	"south"
Dialogue	2	`cuisine`	*international*	"international"
Dialogue	3	`cuisine`	*indian*	"Indian"
Dialogue	4	`number_of_people`	*4*	"4 people"
Dialogue	5	`time_of_booking`	*T16:15*	"16:15"
Dialogue	6	`day_of_booking`	*wednesday*	"wednesday"
Dialogue	7	`time_of_booking`	*T15:15*	"15:15"

(M = 9.16 turns/dialogue, SD = 2.69) generally are longer than the dialogues in CROWDSS (M = 7.94 turns/dialogue, SD = 2.14). For better comparability, we required the random sample[3] to consist of similar-length dialogues. The random sample had a mean dialogue length of 8.25 (SD = 2.16), which, as an unpaired t-test revealed, did not differ significantly from the mean dialogue length in CROWDSS (t = 1.08, p = 0.28). Next, we pre-processed all dialogue turns in both datasets, removing HTML-tags, punctuation and custom stop words (articles, pronouns, prepositions; ≈140 English stop words and ≈320 German stop words with all different case endings, where applicable), as well as tokenizing and lemmatizing the dialogues (relying on spaCy; Honnibal et al., 2020).

Scripting

Our first analysis was aimed at assessing how much freedom the participants had in terms of *what* they should accomplish in the dialogue. We thus compared the entities mentioned in a dialogue's user turns with the entities mentioned in the corresponding scenario (we excluded assistant turns because only users could potentially be influenced by the entities in the scenario). Entities are critical for the tasks of *finding* and *booking a restaurant* as they define key criteria like *cuisine* or *time of booking*. More specifically, we computed the overlap of entity categories (e.g., *Italian, American* for the entity `cuisine`) between dialogue and scenario as a proxy for how pre-*script*-ed a dialogue is: if an entity category in a user turn also appears in the scenario, this counts as overlap. For the MultiWOZ sample, on average 95% of the entity categories mentioned in the user turns in a dialogue also appear in the corresponding scenario for that dialogue (M = 0.95, SD = 0.10). In comparison, in our data, only 75% of entity categories in the user turns also appear in the corresponding scenario (M = 0.75, SD = 0.19). An unpaired t-test revealed that this difference is significant (t = 10.39, p < 0.001). The example dialogue from MultiWOZ in **Table 2** mirrors this result: all entity categories brought up in the user turns also appear in the corresponding scenario. The example dialogue from CROWDSS (**Table 3**), then, shows higher entity category overlap than average, but it points to another interesting aspect: the user asks for a restaurant "[i]n der Nordstadt" (*in the north of town*), even though location is not specified in the corresponding scenario. The map of "Bornberg" (see **Figure 5**) indicates the user's car location with a pin (at the border of the north and east part of town), and this may have led the user to pick *Nordstadt* as a location. The map, besides providing common ground for user and assistant, provides an additional (visual) source of information which is grounded in the task. The map can be used by the users to make an independent choice (picking *Nordstadt* or another part of town) and thus contributes to reducing scripting. Textual elements in the map (the label *Nordstadt*) may have primed the user's lexical choice, but this can easily be avoided, for example, by using a compass instead.

The differing levels of entity category overlap indicate that the user-assistant interactions in the MultiWOZ sample are almost entirely "limited by the complexity of the script" (de Vries et al., 2020; *script* being synonymous with *scenario*), and that the template-based *script* kept the users from making any independent choices which could have led to more diverse dialogues. In CROWDSS, on the other hand, a quarter of all entity categories in the user turns is not *scripted* by the scenario and, thus, they represent independent choices by the users. It seems that our efforts — intentionally not specifying all entities in the scenario and explicitly granting the users some freedom to make their own choices (see Materials and Equipment) — led to relatively diverse dialogues on the level of entity categories.

Second, we had a closer look at dialogues where the same entity appears twice, but in different categories. This would typically happen when the first choice of the user was not available, and an alternative was requested. Such cases should appear in both datasets: in CROWDSS, because we intentionally created scenarios that would lead to zero entries in the assistants' database, making a change of entity category necessary for the dialogue to continue (see Materials and Equipment); and in the

[3]Find the dialogue ids of the MultiWOZ sample in **Supplementary Material**.

TABLE 3 | Entities in scenario and dialogue of CROWDSS example dialogue.

Location	Order	Entity	Category	Surface form
Scenario	1	number_of_people	*2*	"Zum Muttertag möchtest du deine Mama zum Essen einladen"
Scenario	2	cuisine	*vegan*	"Ihr seid beide sehr naturverbunden und esst keine tierischen Lebensmittel"
Scenario	3	garden/terrace	*TRUE*	"Da das Wetter schön sein soll, möchtet ihr gerne draußen sitzen können"
Scenario	4	price_range	*cheap*	"Du befindest dich gerade auf einer finanziellen Durststrecke und hast nur ein begrenztes Budget"
Scenario	5	day_of_booking	*tomorrow*	"morgen"
Scenario	6	time_of_booking	*noon*	"Mittagessen"
Scenario	7	address_req	*address*	"Adresse"
Dialogue	1	price_range	*cheap*	"günstiges"
Dialogue	2	cuisine	*vegan*	"veganes"
Dialogue	3	location	*north*	"In der Nordstadt"
Dialogue	4	garden/terrace	*TRUE*	"mit Garten"
Dialogue	5	number_of_people	*2*	"2"
Dialogue	6	day_of_booking	*tomorrow*	"morgen"
Dialogue	7	time_of_booking	*noon*	"Mittag"

MultiWOZ sample, because there are scenarios following an if-then-logic (e.g., from the scenario in **Table 2**: "The restaurant [. . .] should serve international food. If there is no such restaurant, how about one that serves indian [sic!] food [. . .] you want to book a table [. . .] at 16:15 on wednesday [sic!]. If the booking fails how about 15:15"). Across all dialogues of the MultiWOZ sample, there are 43 entities which appear twice, but in different categories, and for all of them not only the category of the first mention, but also the category of the second mention appears in the scenario. Thus, not only the *first choice* of the user has been scripted, but also the *alternative* that they request when the first choice is unavailable. In comparison, in CROWDSS, out of a total of 31 entities appearing twice in a dialogue, but with different categories, only in four cases both the category of the first mention and second mention appear in the corresponding scenario. These are cases where location is mentioned twice in both the scenario (e.g., as "in deiner Nähe im Osten der Stadt" [*close to your location in the east of town*]) and the dialogue, where the user would first ask for a restaurant *in the east* and then one *nearby* (or vice versa). Crucially, though, in 27 cases the users mention the second entity in a category that is not specified in the scenario. While it is not a surprise that in the MultiWOZ sample all second entity mentions are s*cripted* with regard to the entity category (as the scenarios explicitly predefine the behavior in case a request fails), it is encouraging that the users in CROWDSS continued the dialogues making their own choices in the face of a failed first request, which again increases the dialogue diversity.

Third and last, we investigated whether the dialogues are *scripted* with regard to the entity order. That is, if the order in which entities are mentioned in the scenario provides a sort of a script which is reproduced in the order of entity mentions in the dialogue. For that, we simply compared the order of entities as they are brought up in the dialogue with the order in which they are mentioned in the corresponding scenario (ignoring entities which are specified in the scenario, but not mentioned in the dialogue). In the MultiWOZ sample, 46 of 113 dialogues exhibit the exact same order of entities between dialogue and scenario. In CROWDSS, this is only true of five out of 113 dialogues. Again, the examples in **Table 2** and **Table 3** show that the entity order is identical between dialogue and scenario in the MultiWOZ example, whereas the entities appear in a very different order in the example from CROWDSS. This result, then, also suggests that our situated scenarios (where some entities had to be deduced from multiple pieces of information spread across the scenario, e.g., in **Table 3** the scenario ended with "buche einen Tisch für Euch", where *Euch* had to be derived from "Zum Muttertag möchtest Du Deine Mama zum Essen einladen", at the very beginning of the scenario) led to more diverse dialogues than the scenarios used in MultiWOZ. An interesting open question is whether people have a natural preference for the order in which entities are mentioned. However, our data may be too small to draw any conclusions about this.

Priming

Our next analysis is aimed at assessing how much freedom the participants had in terms of *how* they lexically accomplished their task in the dialogue. For that, we looked at the lexical overlap between a given scenario and the corresponding user turns (we again excluded assistant turns because only the users could potentially be primed by the scenario). Let s be the scenario, u the user turns in a dialogue, and L_s and L_u the set of content word types in s and u, we computed the *lexical overlap* between s and u as

$$LexOverlap_{su} = \frac{|L_s \cup L_u|}{|L_s \cup L_u|}$$

that is, the proportion of content word types in a scenario which are re-used in the user turns. For the MultiWOZ sample, on average more than 50% of the scenario's content word types also appear in the user turns of the corresponding dialogue ($M = 0.51$, $SD = 0.11$). In comparison, in CROWDSS on average only 15% of content word types in the scenario also appear in the user turns ($M = 0.15$, $SD = 0.06$). An unpaired t-test revealed that this difference is significant ($t = 30.29$, $p < 0.001$).

Second, as most of our efforts were focused on how to phrase the *entities* in the scenario, in addition to the lexical priming analysis on the whole content vocabulary, we also compared the lexical overlap for entities only. Compared to our previous analysis of entities, we are now not looking at the normalized

categories but at the overlap between entity surface forms in the user turns with entity surface forms in the corresponding scenario. If an entity surface form in a user turn also appears in the exact same form in the scenario, this counts as overlap. For the MultiWOZ sample, on average more than 84% of the user turn entity surface forms also appear in the corresponding scenario ($M = 0.85$, $SD = 0.16$). In comparison, in CROWDSS on average only 15% of user turn entity surface forms also appear in the scenario ($M = 0.15$, $SD = 0.15$). An unpaired t-test revealed that this difference is significant ($t = 33.27$, $p < 0.001$). The examples in **Table 2** and **Table 3** nicely illustrate that: The surface forms between user turns and corresponding scenario in the MultiWOZ example overlap in their entirety. In the example from CROWDSS, on the other hand, only "morgen" (*tomorrow*) for the day of booking is mentioned in the exact same form in the corresponding scenario. All the other entities take on different surface forms, mainly thanks to the circumlocutions used in the scenario.

Both looking at the whole content vocabulary and looking at entities only, the results suggest that our scenarios led to considerably less priming in the dialogues, when compared to the MultiWOZ sample. Budzianowski et al. (2018) argue that their template-based approach provides "easy-to-follow goals for the users [which] resulted in a bigger diversity and semantical richness of the collected data". However, contrary to this claim, our analysis on the sample appears to confirm the observation by de Vries et al. (2020) that in the worst case the user participants in MultiWOZ seem to have copy-pasted parts of the scenario into their utterances, and in the best case were still "heavily influenced" by the scenario, leading to "unnatural" requests. In CROWDSS, on the other hand, it appears that our relatively small efforts to create "priming-reduced" scenarios encouraged the user to formulate their utterances using their own vocabulary, generating more diverse and natural data.

Scalability

The preceding analyses suggest that our scenarios led to good-quality data. However, handcrafting scenarios like we did is more time-consuming than adopting a template-based approach, even if it arguably leads to more ecologically-valid data. As we were not only aiming at good-quality data, but were also interested in a low-resource approach, we only designed a small set of ten scenarios, which we used as seeds to elicit dialogues by assigning them to participants in a one-to-many ratio. In order to evaluate this tradeoff between good-quality data and low-resource approach, we investigated whether using the same seed still led to diverse dialogue continuations. If this is the case, it would speak for the scalability of our approach, as one could invest time in handcrafting a small set of situated scenarios and elicit diverse dialogues from the same scenario.

In our case, we not only assigned the scenario seeds to different participants, but also assigned dialogue beginnings after the first and in some cases after the second turn to more than one participant as seeds (see Batch-wise data collection). Specifically, we applied a one-to-three participant assignment ratio for collecting batch 1 and 2, and a mixed one-to-one (53 dialogues)/one-to-two ratio (resulting in 30 pairs with the same second turn) for batch 3. In our analysis we look at how diverse dialogues sharing an identical beginning turn out to be from the point where they continue on their own branch (as an example, in **Figure 6**, the dialogues 1.1.1, 1.1.2 and 1.1.3 all share the same scenario and first turn and continue on their own branch from the second turn on). We identified 83 dialogues stemming from 30 dialogue beginnings (unique combinations of scenario + first turn) in order to compare the continuations at two points: 1) at turn/batch 2, and 2) at the end of the dialogues. 53 of these 83 dialogues continued on their own branch at turn 2 (see above). The remaining 30 were selected by randomly choosing one from each of the 30 dialogue pairs which continued on their own branch only at turn 3.

We analyzed each set of dialogues sharing an identical beginning both on the level of DA sequences and vocabulary. A low degree of DA sequence overlap would mean that dialogues with an identical beginning differed substantially regarding *what* participants wanted to accomplish with their utterances. Similarly, a low degree of vocabulary overlap within one set would mean that *how* (i.e., with what words) the participants went about accomplishing their task varied, as they used different words to continue these dialogues, despite their identical beginnings.

For the DA comparison, we relied on Python's built-in SequenceMatcher which looks for the longest contiguous sequences of elements and computes a measure of similarity ranging from 0 to 1. Consider the two second turns which were elicited using the same dialogue beginnings in **Figure 12**. Here, the sequence similarity at the level of the second turn would be 0.5.

We computed the similarity of DA sequences for all possible pairs of dialogues within one set and averaged these similarities over the number of pairs within that set. For the vocabulary comparison, we computed the lexical overlap (see above) between any dialogue pair in the set as the ratio of overlapping word types over all word types in the pair and then averaged over all possible pairs within that set.

Looking only at the second turn, there is a mean similarity of DA sequences of 41% ($M = 0.41$, $SD = 0.28$) and a mean lexical overlap of 32% ($M = 0.32$, $SD = 0.15$). Thus, both on the level of DAs and vocabulary, these second turns *turned out* to be rather diverse, despite their identical dialogue beginnings. Looking at all the turns, there is a mean similarity of DA sequences of 45% ($M = 0.45$, $SD = 0.14$) and a mean lexical overlap of 40% ($M = 0.40$, $SD = 0.07$). Thus, the diversity found for the second turns is maintained until the end of the dialogues. There is a slight decrease in diversity (i.e., an increase in similarity/overlap), but that is to be expected given that there is a finite set of DAs and relevant word types in the given context. The decrease in vocabulary diversity can further be explained by the fact that things which are not addressed in the second turn in one dialogue of a given set may be addressed in a later turn in a different dialogue of the same set. In that case, however, things are done in a different order, which increases the diversity among dialogues on a structural level.

DISCUSSION

Below, we discuss our three goals *good-quality data*, *low-resource approach*, *feasibility in languages with limited crowd-worker availability*.

Good-Quality Data

Looking at scripting and priming, CROWDSS seems to be of higher quality than the MultiWOZ sample. Yet, MultiWOZ is a much larger dataset and we cannot exclude the possibility that the random sample we used is not representative of the dataset as a whole. Our analyses are, however, in line with de Vries et al.'s (2020) observations about MultiWOZ. While CROWDSS does not exhibit any of the other deviations from an ideal, ecologically-valid data collection methodology as listed in de Vries et al. (2020; see above), and despite the encouraging results for scripting and priming, our dataset is in no way perfectly representative of human-machine interaction either. Especially the fact that we collected written data for the SA domain is disadvantageous. As explained, it is, however, not meaningful to collect spoken data in an asynchronous fashion. To collect spoken data, a live interaction on a dedicated platform would have been necessary, but that did not fit in with our second and third goal (*low-resource, feasibility in languages with limited crowd-worker availability*). Therefore, we restricted ourselves to putting emphasis on the fact that the dialogues are a simulation of spoken SA interactions (see Materials and Equipment). The mean turn length in tokens is relatively short in CROWDSS ($M = 8.4$, $SD = 1.7$, compared to $M = 11.46$, $SD = 2.37$ in the MultiWOZ sample), which is encouraging considering the fact that spoken dialogue typically consists of shorter utterances compared with written interaction. Hauptmann and Rudnicky (1988) report an average command turn length of 6.1 tokens for a WOz setup in the spoken modality (speaking to a computer/wizard; see also Fraser and Gilbert 1991). In sum, our investments into reducing scripting and priming seem to have paid off.

Low-Resource Approach

Our second goal was to collect dialogue data with as little resources as possible. For that reason, we collected data in an asynchronous way, which essentially obviated the need for a technically more complex setup that would enable live interactions. The asynchronous setup entailed that we collected written data, which was again low-resource because participants only needed a screen and a keyboard. Microphones and an ASR module on our side were not necessary. Relying only on AMT's HTML editor naturally constrained our design possibilities, but it is important to note that we were able to design fully functional and aesthetically appealing interfaces. This made it possible to collect all data on-site, rather than having to host the data collection on an external website, which would have required more resources.

Thus, it turned out to be feasible to reduce the technical overhead and only rely on an existing platform for data collection. At the same time, there was considerable manual overhead, mostly in between the batches for running a post-processing script on the most recent batch, reviewing flagged dialogues, excluding poorly performing participants, accepting the HITs so that the participants would receive payment, and preparing the following batch. Most of these steps can be automated using scripts, and the manual review in between batches can be skipped or downscaled depending on where the compromise between quality and quantity should be made.

An encouraging finding regarding the manual overhead is that our resource-friendly approach to only handcraft ten scenarios and use them to elicit more than the tenfold of dialogues still appears to have led to diverse dialogues. Therefore, we argue that the investments into good-quality data that we propose, namely the situated scenarios, are also implementable in large-scale dialogue data collections, as it does not seem necessary to have one unique scenario per dialogue. Thus, it appears that collecting good-quality data can go hand in hand with a low-resource setup, which is a step towards reconciling quality with quantity.

Feasibility in Languages With Limited Crowd-Worker Availability

Our third goal was to make sure that our data collection approach worked for German, where fewer crowd-workers are available compared to English. This proved to work well for our data collection. We see two main reasons for that. First, thanks to the batch-wise setup, one participant could contribute to multiple dialogues, though always in the same role. Using sub-batches we could still ensure diversity in participants which, if the data is used for training an algorithm, could allow the model to generalize better (Geva et al., 2019). Second, the asynchronous setup obviated the pressure to pair up two simultaneously available crowd-workers for a live interaction. Instead, we could launch a batch and simply wait until all dialogues were continued which would typically take a couple of hours.

Since we only collected a small dataset, we did not test the boundaries of German-speaking crowd-worker availability and we can therefore not say how large the German worker pool is. However, as NLP datasets are needed for languages other than English where they often are collected in ways that are not feasible in smaller languages (e.g., live interactions) we argue that our approach is a step in the right direction, enabling dialogue data collections for different languages than English.

Future Research

The main limitation of this work is the small size of the dataset. Thus, a larger dataset collected with our situated scenarios may be needed to strengthen the generalizability of the analyses reported above. Another limitation is the written modality which we had to rely on for resource and crowd-worker availability reasons. A next step could be to compare our dataset to a corpus of spoken SA interactions (e.g., Siegert, 2020) in order to evaluate potential modality-induced differences. Speaking of modalities, it would also be interesting to include more visual elements like the map in the scenarios, which could enhance the situatedness of the task and further reduce scripting and priming. Lastly, it could be worth combining a low-resource template-based approach with our suggestions concerning vague language and circumlocutions. One could use *whole* sentences formulated along these lines to build scenarios, rather than just inserting one-word entities in designated placeholders in otherwise ready-made scenarios, as is the case with the traditional template-based approach.

CONCLUDING REMARKS

In this paper, we have presented a novel approach for dialogue data collection. We argue that our framework makes it possible to crowdsource good-quality dialogues in a fairly low-resource fashion which is furthermore feasible in other languages than English. Additionally, we show that putting little effort in creating instructions and scenarios for the participants can lead to better-quality data than a template-based approach like in MultiWOZ. Our comparison with a sample from MultiWOZ suggests that our endeavors led to significantly less scripting and priming and thus more ecologically-valid dialogue data.

With the shift from rule-based to machine learning-based NLP systems, recently, datasets have focused heavily on quantity, under the assumption that the mass of data still comes in at least "appropriate" quality (de Vries et al., 2020). While quantity will always come at the cost of quality, we argue that small investments in the collection setup such as the ones we propose, and which are also possible for large-scale collections, can go a long way in improving data quality. WOz is a very helpful tool to "bootstrap out of [the] chicken and egg problem" (de Vries et al., 2020), that is, the problem that we do not know how humans would talk to human-like machines if they existed, however to make them come into existence we need data from this type of interaction. We argue that if one puts time and effort into a WOz collection, it is worth to also invest in reducing researcher bias and instead allowing the participants to interact as naturally as possible in the given context. Taking it to the extreme, you could even wonder what WOz data can give you beyond a machine-generated "dialogue" if you do not afford participants the opportunity to phrase their utterances in a natural way.

DATA AVAILABILITY STATEMENT

The dataset presented in this study can be found in an online repository. The name of the repository and accession number can be found here: https://dx.doi.org/10.24406/fordatis/124.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

YF and AZ contributed to conception and design of the data collection. YF managed the data collection process and performed data quality analyses. YF wrote the first draft of the manuscript. YF and AZ wrote sections of the manuscript. YF and AZ contributed to manuscript revision, read, and approved the submitted version.

ACKNOWLEDGMENTS

The authors would like to acknowledge Anna Sauer, Valmir Kosumi and Zahra Kolagar for their contribution to the data collection and analysis as well as Junbo Huang for his feedback on the article. We acknowledge the support of the Federal Ministry for Economic Affairs and Energy (BMWi) project SPEAKER (FKZ 01MK20011A).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomp.2021.686050/full#supplementary-material

REFERENCES

Asri, L. E. l., Schulz, H., Sharma, S., Zumer, J., Harris, J., Fine, E., et al. (2017). "Frames: A Corpus for Adding Memory to Goal-Oriented Dialogue Systems," in Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue, Saarbrücken, Germany, August 2017, 207–219. Available at: https://www.aclweb.org/anthology/W17-5526

Barrus, T. (2019). *Pure Python Spell Checker Based on Work by Peter Norvig*. Available at: pypi.org/project/pyspellchecker/.

Barsalou, L. W. (2009). Simulation, Situated Conceptualization, and Prediction. *Phil. Trans. R. Soc. B* 364 (1521), 1281–1289. doi:10.1098/rstb.2008.0319

Bocklisch, T., Faulkner, J., Pawlowski, N., and Nichol, A. (2017). Rasa: Open Source Language Understanding and Dialogue Management. *ArXiv*, 1–9. http://arxiv.org/abs/1712.05181.

Branigan, H. P., Catchpole, C. M., and Pickering, M. J. (2011). What Makes Dialogues Easy to Understand? *Lang. Cogn. Process.* 26 (10), 1667–1686. doi:10.1080/01690965.2010.524765

Budzianowski, P., Wen, T.-H., Tseng, B.-H., Casanueva, I., Ultes, S., Ramadan, O., and Gašić, M. (2018). "MultiWoz - A Large-Scale Multi-Domain Wizard-Of-Oz Dataset for Task-Oriented Dialogue Modelling," in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP, Brussels, Belgium, October–November 2018, 5016–5026. doi:10.18653/v1/d18-1547

Buhrmester, M., Kwang, T., and Gosling, S. D. (2011). Amazon's Mechanical Turk. *Perspect. Psychol. Sci.* 6 (1), 3–5. doi:10.1177/1745691610393980

Bunt, H., Jan, A., Jean, C., Choe, J. W., Alex, C. F., Hasida, K., et al. (2010). "Towards an ISO Standard for Dialogue Act Annotation," in Proceedings of the 7th International Conference on Language Resources and Evaluation, LREC, Valletta, Malta, May 2010, 2548–2555.

Byrne, B., Krishnamoorthi, K., Sankar, C., Neelakantan, A., Goodrich, B., Duckworth, D., et al. (2019). Taskmaster-1: Toward a Realistic and Diverse Dialog Dataset. Proceedings of the Conference, 4516–4525. doi:10.18653/v1/d19-1459

Chen, B. X., and Metz., C. (2019). "Google's Duplex Uses A.I. To Mimic Humans (Sometimes). *New York Times*. May 22, 2019. Available at: https://www.nytimes.com/2019/05/22/technology/personaltech/ai-google-duplex.html.

Clark, H. H. (1996). *Using Language*. Cambridge, United Kingdom: Cambridge University Press. doi:10.1017/CBO9780511620539

Clark, H. H., and Wilkes-Gibbs., D. (1986). Referring as a Collaborative Process. *Cognition* 22 (1), 1–39. doi:10.1016/0010-0277(86)90010-7

de Vries, H., Bahdanau, D., and Manning, C. (2020). Towards Ecologically Valid Research on Language User Interfaces. ArXiv http://arxiv.org/abs/2007.14435.

de Visser, E. J., Monfort, S. S., Ryan, M. K., Melissa, A., Smith, B., McKnight, P. E., et al. (2016). Almost Human: Anthropomorphism Increases Trust Resilience in Cognitive Agents. *J. Exp. Psychol. Appl.* 22 (3), 331–349. doi:10.1037/xap0000092

Dürscheid, C., and Brommer, S. (2009). Getippte Dialoge in Neuen Medien. Sprachkritische Aspekte Und Linguistische Analysen. *Linguistik Online* 1 (37), 3–20. Available at: https://www.zora.uzh.ch/id/eprint/24486/.

Eric, M., Krishnan, L., Charette, F., and Manning, C. D. M. (2017). "Key-Value Retrieval Networks for Task-Oriented Dialogue," in Proceedings of the Conference, Saarbrücken, Germany, August 2017, 37–49. doi:10.18653/v1/w17-5506

Fox, Tree., and Jean, E. (1999). Listening in on Monologues and Dialogues. *Discourse Process.* 27 (1), 35–53. doi:10.1080/01638539909545049

Fraser, N. M., and Gilbert, G. N. (1991). Simulating Speech Systems. *Comput. Speech Lang.* 5 (1), 81–99. doi:10.1016/0885-2308(91)90019-m

Garcia, Chiyah., Francisco, J., Lopes, José., Liu, X., and Hastie, H. (2020). "CRWIZ: A Framework for Crowdsourcing Real-Time Wizard-Of-Oz Dialogues," in Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, France, May 2020, 288–297. Available at: https://www.aclweb.org/anthology/2020.lrec-1.36

Garrod, S., and Anderson, A. (1987). Saying what You Mean in Dialogue: A Study in Conceptual and Semantic Co-ordination. *Cognition* 27 (2), 181–218. doi:10.1016/0010-0277(87)90018-7

Geva, M., Goldberg, Y., and Berant, J. (2019). "Are We Modeling the Task or the Annotator? an Investigation of Annotator Bias in Natural Language Understanding Datasets," in EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference, Hong Kong, China, November 2019 (Association for Computational Linguistics). 1161–1166. doi:10.18653/v1/d19-1107

Grosz, B. J. (2018). Smart Enough to Talk with Us? Foundations and Challenges for Dialogue Capable AI Systems. *Comput. Linguistics* 44 (1), 1–15. doi:10.1162/COLI_a_00313

Hauptmann, A. G., and Rudnicky, A. I. (1988). Talking to Computers: An Empirical Investigation. *Int. J. Man-Machine Stud.* 28 (6), 583–604. doi:10.1016/S0020-7373(88)80062-2

Henderson, M., Thomson, B., and Williams, J. (2013). Dialog State Tracking Challenge 2 & 3." Dialogues with Social Robots, No. September: 1–22. Available at: http://camdial.org/~mh521/dstc/downloads/handbook.pdf.

Honnibal, M., Montani, I., Van Landeghem, S., and Boyd, A. (2020). SpaCy: Industrial-Strength Natural Language Processing in Python. Zenodo. doi:10.5281/zenodo.1212303

Jonell, P., Fallgren, P., Doğan, F. I., Lopes, J., Irmak Doğan, F., Wennberg, U., et al. (2019). Crowdsourcing a Self-Evolving Dialog Graph. CUI'19: Proceedings of the 1st International Conference on Conversational User Interfaces, Dublin, Ireland, August 2019. doi:10.1145/3342775.3342790

Kelley, J. F. (1983). "An Empirical Methodology for Writing User-Friendly Natural Language Computer Applications," in CHI '83: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, New York, NY, December 1983, 193–196.

Levelt, W. J. M., and Kelter., S. (1982). Surface Form and Memory in Question Answering. *Cogn. Psychol.* 14 (1). doi:10.1016/0010-0285(82)90005-6

Leviathan, Yaniv., and Matias, Yossi. (2018). Google Duplex: An AI System for Accomplishing Real-World Tasks over the Phone. Available at: https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html.

McCarthy, P. M., and Jarvis., S. (2010). MTLD, Vocd-D, and HD-D: A Validation Study of Sophisticated Approaches to Lexical Diversity Assessment. *Behav. Res. Methods* 42 (2), 381–392. doi:10.3758/BRM.42.2.381

McRae, K., Nedjadrasul, D., Pau, R., Bethany, P. H. L., Lo, B. P.-H., and King, L. (2018). Abstract Concepts and Pictures of Real-World Situations Activate One Another. *Top. Cogn. Sci.* 10 (3), 518–532. doi:10.1111/tops.12328

Merdivan, E., Singh, D., Hanke, S., Kropf, J., Holzinger, A., and Geist, M. (2020). Human Annotated Dialogues Dataset for Natural Conversational Agents. *Appl. Sci.* 10 (762), 762. doi:10.3390/app10030762

Nakayama, H., Kubo, T., Kamura, J., Taniguchi, Y., and Liang, X. (2018). Doccano: Text Annotation Tool for Human. https://github.com/doccano/doccano.

Nass, C., and Lee, K. M. (2001). Does Computer-Synthesized Speech Manifest Personality? Experimental Tests of Recognition, Similarity-Attraction, and Consistency-Attraction. *J. Exp. Psychol. Appl.* 7 (3), 171–181. doi:10.1037//1076-898X.7.3.17110.1037/1076-898x.7.3.171Lee

Nass, C., Steuer, J., and Tauber, E. R. T. (1994). "Computers Are Social Actors," in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'94) (New York, NY: Association for Computing Machinery), 72–78. doi:10.1145/191666.191703

Nishimaki, K. (2014). Characteristics of Spoken and Written Communication in the Opening and Closing Sections of Instant Messaging. Dissertations and Theses. Paper 1548. Available at: https://pdxscholar.library.pdx.edu/open_access_etds.

Oreström, B. (1983). *Turn-Taking in English Conversation*. Lund: Krieger Publishing Company.

Pareti, Silvia., and Lando, Tatiana. (2019). "Dialog Intent Structure: A Hierarchical Schema of Linked Dialog Acts," in LREC 2018 - 11th International Conference on Language Resources and Evaluation, Miyazaki, Japan, May 2018, 2907–2914. Available at: https://www.aclweb.org/anthology/L18-1460/

Pickering, M. J., and Garrod., S. (2004). Toward a Mechanistic Psychology of Dialogue. *Behav. Brain Sci.* 27 (2). doi:10.1017/s0140525x04000056

Poesio, M., and Rieser, H. (2011). An Incremental Model of Anaphora and Reference Resolution Based on Resource Situations. *dad* 2 (1), 235–277. doi:10.5087/dad.2011.110

Rastogi, A., Zang, X., Sunkara, S., Gupta, R., and Khaitan, P. (2020). "Towards Scalable Multi-Domain Conversational Agents: The Schema-Guided Dialogue Dataset," in Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, February 2020. 34, 8696–96. doi:10.1609/aaai.v34i05.6394

Reeves, B., and Nass, C. (1996). *Media Equation: How People Treat Computers, Television, and New Media like Real People and Places*. Stanford, CA: The Center for the Study of Language and Information Publications. doi:10.1300/j105v24n03_14

Rieser, V. (2008). *Bootstrapping Reinforcement Learning-Based Dialogue Strategies from Wizard-Of-Oz Data*." Saarbrücken, Germany: German Research Center for Artificial Intelligence.

Rieser, V., and Lemon, O. (2011). "Reinforcement Learning for Adaptive Dialogue Systems. A Data-Driven Methodology for Dialogue Management and Natural Language Generation," in *Theory and Applications of Natural Language Processing*. Editors G Hirst, E Hovy, and M Johnson (Heidelberg, Dordrecht, London, New York: Springer). doi:10.1007/978-3-642-24942-6http://www.springer.com/series/8899

Schegloff, E. A. (1982). *"Discourse as an Interactional Achievement: Some Uses of 'uh Huh' and Other Things that Come between Sentences*. Washington, DC: Roundtable on Languages and Linguistics.

Schlangen, D. (2019). Language Tasks and Language Games: On Methodology in Current Natural Language Processing Research. ArXiv https://arxiv.org/abs/1908.10747. doi:10.18653/v1/w19-0424

Schnoebelen, T., and Kuperman, V. (2010). Using Amazon Mechanical Turk for Linguistic Research. *Psihologija* 43 (4), 441–464. doi:10.2298/PSI1004441S

Serban, I. V., Lowe, R., Henderson, P., Charlin, L., and Pineau, J. (2018). A Survey of Available Corpora for Building Data-Driven Dialogue Systems: The Journal Version. *dad* 9 (1), 1–49. doi:10.5087/dad.2018.101

Shah, P., Hakkani-Tür, D., Tür, G., Rastogi, A., Bapna, A., Nayak, N., and Heck, L. (2018). "Building a Conversational Agent Overnight with Dialogue Self-Play," in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, LA, June 2018 (Association for Computational Linguistics), Volume 3, 41–51. (Industry Papers) https://www.aclweb.org/anthology/N18-3006.pdf.

Shatz, M., and Gelman, R. (1973). The Development of Communication Skills: Modifications in the Speech of Young Children as a Function of Listener. *Monogr. Soc. Res. Child Develop.* 38 (5), 1. doi:10.2307/1165783

Siegert, I. (2020). in Alexa in the Wild' - Collecting Unconstrained Conversations with a Modern Voice Assistant in a Public Environment." LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings, Marseille, France, May 2020, 615–619. https://www.aclweb.org/anthology/2020.lrec-1.77/.

Stalnaker, R. (1978). Assertion. *Syntax and Semantics* 9, 315–332.

Wang, W. Y., Bohus, D., Kamar, E., and Horvitz, E. (2012). "Crowdsourcing the Acquisition of Natural Language Corpora: Methods and Observations. "Crowdsourcing the Acquisition of Natural Language Corpora: Methods and

Observations.," in 2012 IEEE Workshop on Spoken Language Technology, SLT 2012 - Proceedings, Miami, FL, December 2012. doi:10.1109/SLT.2012.6424200

Wei, W., Le, Q., Dai, A., and Li, J. (2018). "Airdialogue: An Environment for Goal-Oriented Dialogue Research," in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP, Brussels, Belgium, October–November 2018. doi:10.18653/v1/d18-1419LeLi2018

Wen, T.-H., Vandyke, D., Mrkšić, N., Gašic, M., Rojas Barahona, L. M., Su, P.-H., Ultes, S., and Young, S. (2017). "A Network-Based End-To-End Trainable Task-Oriented Dialogue System," in 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017 - Proceedings of Conference, Valencia, Spain, April 2017, 1, 438–449. doi:10.18653/v1/e17-1042

Williams, J. D., and Young, S. (2007). Partially Observable Markov Decision Processes for Spoken Dialog Systems. *Comput. Speech Lang.* 21 (2), 393–422. doi:10.1016/j.csl.2006.06.008

Xu, Y., and Reitter, D. (2018). Information Density Converges in Dialogue: Towards an Information-Theoretic Model. *Cognition* 170, 147–163. doi:10.1016/j.cognition.2017.09.018

Yuan, S., Brüggemeier, B., Hillmann, S., and Michael, T. (2020). "User Preference and Categories for Error Responses in Conversational User Interfaces," in CUI'20: Proceedings of the 2nd Conference on Conversational User Interfaces, Bilbao, Spain, July 2020. doi:10.1145/3405755.3406126

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.