RoboClean: Contextual Language Grounding for Human-Robot Interactions in Specialised Low-Resource Environments

Carolina Fuentes
School of Computer Science and Informatics
Cardiff University
Cardiff, UK
FuentesToroC@cardiff.ac.uk

Martin Porcheron
Bold Insight
London, UK
martin@boldinsight.co.uk

Joel E. Fischer
Mixed Reality Laboratory, School of Computer Science
University of Nottingham
Nottingham, UK
joel.fischer@nottingham.ac.uk

ABSTRACT

Building effective voice interfaces for the instruction of service robots in specialised environments is difficult due to the local knowledge of workers, such as specific terminology for objects and space, leading to limited data to train language models (known as 'low-resource' domains) and challenges in language grounding. We present a language grounding study in which we a) elicit spoken natural language of context experts *in situ* through a Wizard of Oz study and compile a dataset, b) qualitatively examine linguistic properties of the resulting instructions to reveal referential categories and parameters employed to construct instructions in context. We discuss how our language grounding protocol may be applied to bootstrap a language model in its targeted use context. Our work contributes a linguistic understanding of robot instructions that can be applied by designers and researchers to develop spoken language understanding for human-robot interactions in specialised, low-resource environments.

CCS CONCEPTS

• **Computer systems organization → Robotic control**; • **Human-centered computing → Natural language interfaces**; *User studies.*

KEYWORDS

human-robot interaction, speech, conversational interfaces, language grounding, HRI, spoken language understanding, SLU

ACM Reference Format:
Carolina Fuentes, Martin Porcheron, and Joel E. Fischer. 2023. RoboClean: Contextual Language Grounding for Human-Robot Interactions in Specialised Low-Resource Environments. In *ACM conference on Conversational User Interfaces (CUI '23), July 19–21, 2023, Eindhoven, Netherlands.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3571884.3597137

1 INTRODUCTION

Enabling humans to use natural language to interact with robots holds many promises, including empowering people with little or no training in complex control interfaces or programming to make use of robots. As robots become more capable to support humans in environments including factories, homes, workplaces and public spaces [28, 59], it may be important to provide simple and intuitive ways to interact with and control them, including in a way that can be understood by other humans in the setting [26].

Although human-human dialogue is shown to have parallels with human-machine dialogue, fundamental differences exist that warrant those building such machines to pay close attention to the design of these interactional technologies [47]. One of the core challenges to enable natural language interaction is posed by the need for language grounding [19, 56]. Language grounding in robotics is needed to map words to aspects of the physical world and physical actions [27]. To enable a robot to carry out an apparently simple instruction such as "put the yellow block into the box on your left" requires natural language and image processing, object detection and intent matching (including physical reference resolution) to ground the semantic representations (e.g., 'yellow block') to the percept of the physical object. Technically speaking, grounding relies either on lexical approaches that formalise the relationship between semantic and physical representations or learning language models from large datasets [56].

In the real world, the language used to describe environments and the objects in them is rich and varied, making it difficult to rely on existing generalist language models. Considering the potential instruction given by a factory worker to 'clean next to the rig'; the token 'rig' may present a previously unseen token for a generalist language model. Thus, these kinds of specialised environments can be characterised as 'low-resource' from a modelling perspective as there is limited data available to train the models [33]. In this article, we address part of the challenge of language grounding for specialised low-resource environments, such as factories, where potential operators possess local knowledge (e.g., about what the objects in the space are called); and this local knowledge shapes the specific terminology operators may use to instruct a robot. As tasks and the environment become more complex, changing one's terminology in order to maintain successful communication with the robots is cumbersome and runs against basic human-centred design principles [7]. Particularly for specialised working environments like laboratories, production, or industrial plants [52], asking workers to alter their behaviour and learn new "rules of engagement" incurs various costs for operators, such as those of time, where it may involve retraining and alteration of established, reasoned working practices. In addition, the results of "misunderstandings" between the robots and the human workers can have much more direct consequences here in terms of worker safety, production timelines, legislation, and industry hygiene standards guidelines (e.g. [4]).

CUI '23, July 19–21, 2023, Eindhoven, Netherlands

We present an elicitation study of natural language instructions in specialised environments through a robotic vacuum cleaner ('robovac') case study (see below) to inform future language grounding models. We conducted a Wizard of Oz (WOz) elicitation study to collect a dataset of people's spoken instructions of robovac in a food chemistry laboratory [40]. Prior work has demonstrated WOz studies as a viable intermediate step in automated robot development, although continued to rely upon human intelligence for landmarks as part of the robot navigation [3]. We then employed a mixed-methods analysis to examine the structure of language by the participants.

We use qualitative linguistic analysis, drawing upon referential language analysis, to "situate the communication in the current task and environmental state" [35] and to study the context and efficacy of instructions. We discuss how the combination of language elicitation *in situ*, mixed methods analysis and Machine Learning could be used by researchers and designers to verify the usefulness of data to create spoken language understanding for specific low-resource environments, enabling development of robots that can be controlled through locally specific natural language.

The primary contributions of this work are (1) the *RoboClean* dataset obtained from the Wizard of Oz elicitation study in a laboratory setting[1], and (2) insights on the linguistic (e.g., referential) make-up of instructions for use in future spoken language understanding (SLU) for human-robot interaction.

1.1 RoboClean: a case study

We present results from a study in a laboratory setting in which participants interacted with a 'robovac'. Robovacs are typically tasked with cleaning unbounded areas, with users issuing simple commands to clean (often, but not exclusively through software or hardware buttons). Typically, the robot has to assess the layout of the environment, formulate and execute actions to clean it, ideally without requiring interruption or human intervention.

Robovacs present what we consider to be an interesting case, as they are widely used in both domestic [14] and specialised environments [49]. We focus on the latter setting, which poses various different challenges and distinctions from domestic settings. The consequences of misunderstanding a command in these settings is more serious than in a domestic environment. Operators in specialised environments have local knowledge such as vernacular terminology to denote object and areas in the space, and an operator may wish to prioritise certain areas, enquire about past tasks, or schedule tasks for the future.

The purpose of the case study is to elicit and examine the language that persons with local knowledge use to instruct the robovac in a number of cleaning scenarios. To do this we conducted a WOz-based elicitation study in a university food chemistry laboratory. We recruited people who regularly work in the laboratory and thus possess (at least some) local knowledge regarding the objects and areas in the lab. The RoboClean case study thus helps us understand what kind of approach might work to cater for other specialised environments with broader robotic instructional needs. We then make use of the resulting dataset to examine the language further to understand the requirements to develop spoken language understanding for human-robot interaction.

This paper is structured as follows: section 2 presents research related to our study, then the approach is introduced in section 3, the qualitative analysis and understanding of spoken instructions are described in section 4, the discussion and design implications are in section 5, and we offer our conclusion in section 6.

2 BACKGROUND

In food factories, cleaning floors is still largely completed by human workers and offers an opportunity to explore human-robot collaboration to assist in this food safety-critical activity. Cleaning 'co-bot' teams have been proposed as a way of integrating robotic cleaning with on-line sensing of allergens using near-infrared spectroscopy (e.g. [5, 46]). Robot-vacuum cleaners have been explored in domestic settings revealing the impact on people's routines, activities, and roles [12, 13]. We expand this work by investigating what kind of contextual language understanding might be needed to ensure effective and situationally appropriate human-robot interaction in industrial settings.

In this section, we review the literature studying interaction with robots through natural language, outline the language grounding problem for specialised, low-resource environments, and relevant HRI datasets.

2.1 Challenges in natural language interaction for HRI

To improve how robots can collaborate with people to engage in complex tasks, one of the key elements is the use of natural language to maximise the robot's verbal and non-verbal understanding [30, 57]. Spoken Dialogue Systems (SDSs) enable interaction by natural language communication, ranging from simple pre-defined Q&A dialogues to more sophisticated interactions by conversational agents. The main SDSs components involve speech recognition, language understanding, dialogue management, communication with external systems, response generation, and speech output [32]. Similar to SDSs, spoken language understanding (SLU), involves automatic speech recognition, natural language processing, understanding, and synthesis [6, 25]. Voice interfaces have followed a more commercial deployment approach [31], like those present in smart speakers, acting as voice agents or virtual assistants and being ever more present in everyday life [41].

From a HRI perspective there is still a limited understanding of how people respond to robots in complex settings and how social dynamics are affected [21]. In terms of deployments of collaborative robots (co-bots) in industrial contexts, Sauppé and Mutlu [51] conclude that co-bots should be designed with basic language capabilities that would allow them to communicate with their human counterparts whenever they malfunction. In contexts, such as tour-guide robots, identifying patterns of pronouns and nouns resulting in accurate responses to visitors' enquiries were identified as important [22]. Both of these studies suggest that the language capabilities and communication patterns of the robot are essential to the design of new interfaces.

[1]The dataset is available under a CC-BY license at https://doi.org/10.17639/nott.7295

For the design of robotic assistants for hospitals, in which robots are responsible for the completion of simple tasks such as interacting with staff and the delivery of messages and medicines, SDSs offer a state-based approach for supporting dialogues by pre-defining dialogue into structures and series of states. This means that whilst the users' inputs are more easily predicted by the robot, the flexibility of the dialogue is considerably limited. The dialogue of the robot is therefore designed to consist of short exchanges, with the main focus being ensuring correct interpretation of the users' inputs. Various observations were made regarding the prospect of mobile robots with natural language interfaces, including that given the multiple subsystems such robots contain (e.g., navigation), SDSs with a state-based dialogue are a viable approach given the relatively fast language processing possible [54].

Tellex et al. [57] evaluate the performance and precision of a model capable of understanding natural language commands from untrained users for instructing an autonomous forklift and suggest a probabilistic model to structure spatial description clauses. However, the latest investigations in this area suggest the need for in-depth contextual understanding to allow utterances to be more effectively understood within specific contexts [29, 56]. Thus, examining robots in different contexts, across different groups, and different types of workplaces, is critical to expanding the required knowledge for fluid human-robot communication, including more effective interactions and task performance [21, 29, 30].

2.2 Natural language grounding in low-resource environments

Natural language present a rich set of challenges for HRI and AI: referential expression resolution, multimodality, the dynamism of language, ambiguities and automatic generation of referring expressions, the seemingly imprecise nature of utterances and 'unstructured' ways for describing objects, situations or directions [10, 15, 17, 23, 30, 50, 55, 57, 60]. These challenges stem from applying computational formalisms to language, rather than the other way around. When humans communicate or collaborate with others they can use different levels of internal and external strategies to deal with language challenges, but for making robots capable of appropriate responses to human requests it is necessary to embed tools, strategies, and data for knowledge that improves the machine reasoning process [50].

Adaptable and situated robots that can understand human practices when interacting via speech are critical for successful language-based human-robot interactions [56]. While we acknowledge that Automatic Speech Recognition (ASR) is necessary to recognise what the human is saying (i.e., machine transcription) [1, 45], our work takes this aspect of the SLU pipeline as a given. Once human speech has been transcribed, in order for the robot to understand what the human may have meant, semantic recognition through natural language processing is required. This is linked to the language grounding problem (related to the symbol grounding problem) [19, 56].

The language/symbol grounding problem, sometimes known as "grounded language understanding" [19, 56] has been extensively studied within robotics. It centres on the grounding of semantic representations (e.g., as a result of language processing) in the physical percepts (e.g., as a result of image processing) and physical actions (e.g., as a result of motion plan formulation) to achieve grounded language acquisition. Grounding relies either on lexical approaches that formalise the relationship between semantic and physical representations or learning language models from large datasets [56]. However, the grounding for specific environments is unlikely to be able to rely on existing generalised datasets due to the specialised local terminology and objects in the space. Specific large training datasets will be costly to generate. Hence, it is necessary to develop models capable of operating under data-constrained environments better known as Low-Resource Environments (LRE) [33, 62]. LREs relate to constrained scenarios where it is difficult to collect large volumes of data needed to train models. Some examples of low-resource environments are specialised technical contexts and their specific languages, conversational robots interacting with older adults with dementia or people with speech impediments [33, 43]. To situate robots that successfully process natural language in these specialised, low-resource scenarios, it will be critical to generate datasets that contain locally specific language so that robots can improve their natural language grounding capabilities.

To our knowledge, there are a few datasets in HRI from low-resource environments, and more are needed to be tested with different algorithms [34] including reusable datasets from small domains [29]. Some types of large datasets have also been found to be not not applicable in low-resource environments [44]. A survey presented in 2020 listed the most common datasets used in language grounding and robots, highlighting the conflict between the data domain and the current robotic task [56]. Moreover, matching these rich datasets and the contextual environment where the robot will operate would be challenging [56].

In this work, we collect our own corpus of data to study participants' natural/unstructured instructions in the context of delegating cleaning tasks to a robot vacuum cleaner. Following, we present our approach to generate a contextualised dataset that contains particular linguistic structures for language grounding in low-resource environments.

3 APPROACH

Our approach consists of an elicitation study followed by linguistic analysis. Similar approaches have been tried previously to develop natural language interfaces (e.g., [24]). As such, we conducted qualitative analysis of the spoken instructions (linguistic analysis), which may be used to inform further work using Machine Learning techniques to develop models based on the language practices identified from *in situ* interactions.

3.1 Data collection and curation

We designed an elicitation study to examine and understand the situated nature in which instructions are made in a 'low-resource' environment in a similar approach to Bonial et al. [3]. As discussed above, we opted to use a robotic vacuum cleaner in a university food chemistry laboratory. We chose this for its specialised industrial nature with the laboratory equipment present, which with its concomitant specialised terminology represents a low-resource environment for spoken reference resolution. This setting also has greater accessibility for us to conduct safe studies in comparison to an always-active industrial environment. Likewise, such vacuums

are a readily available commercial technology that could be purposed for our needs with relative ease. We hasten to add that we do not proffer the data though this elicitation study as applicable to all similar settings but rather as a singular case study. Thus, we seek to contribute a generalisable approach that can be applied in other low-resource environments rather than a single bespoke dataset.

We now introduce the design of our study and the curated scenarios we asked participants to complete, the details of the Wizard of Oz methodology, and our analytic approach. Our study was approved by the School of Computer Science Research Ethics Committee and each participant completed an informed consent process prior to the commencement of the study. Each participant received a £20 gift voucher for their time.

3.1.1 Study design. To explore how people issue commands to the robot we recruited 21 participants to complete a research study in the laboratory. Participants instructed the robot to clean areas of the laboratory floor. Figure 1 shows the starting position of robots within the room (R1 and R2) as well as existing objects in the space (tanks, tables, and a bleed-gauge, labelled as *O1–O6*). We defined three arbitrary areas (*A1*, *A2*, and *A3*) in the laboratory and participants were asked to send robot to clean any of these. Neither the objects nor the areas were given specific *a priori* names, rather our intention was to collect participants' natural descriptions (references) of these, from which we could develop our protocol. As the intention of the protocol was to enable rapid deployments of existing technologies to low-resource settings, these predefined areas act as constraints, built upon *assumptions* about what practices might be needed. Participants' response to these assumptions will in turn shape our understanding of needs of the technology for these settings. This stands in comparison to elicitation studies that are more exploratory in nature, focusing on informing the design of a technology rather than rapid deployments [20].

We use a Wizard of Oz (WOz) approach [11, 40, 48, 53] rather than implementing the 'intelligence' of the robot's spoken language understanding and control ourselves. In this approach, a second researcher (the *Wizard*) controlled the robot's movements and audible response without the participant's prior knowledge. The Wizard sat at the table to the left of the laboratory (see Figure 1) and was introduced to participants simply as another researcher observing the study—they never interacted with the participants and the Wizard-controlled nature of the robots were not explained to participants until after the study. We made use of two Neato D7 Robot Vacuum Cleaners [36, 58] as our robots, controllable using the Neato mobile app installed on an iPad. Although participants were only ever 'controlling' one robot at a time (the other was a *hot spare*). Once the robot vacuum cleaner started to clean, it worked autonomously within the zone to be cleaned, but would not detect debris on the floor, instead following the internal programme for cleaning.

We used existing software for running voice-based Wizard of Oz studies [42]. The software allows for the partial pre-scripting of responses and is designed to reduce the amount of typing required in generating responses. The system also logs data, such as timestamps and messages spoken to support later analysis. For this study, the software piped the typed output to the macOS voice subsystem—specifically using the voice 'Daniel' (a British English male voice). The computer running the software was connected via Bluetooth to a portable battery-powered speaker attached to the top of the robot. This furthers the simulation of the user interacting with the robot.

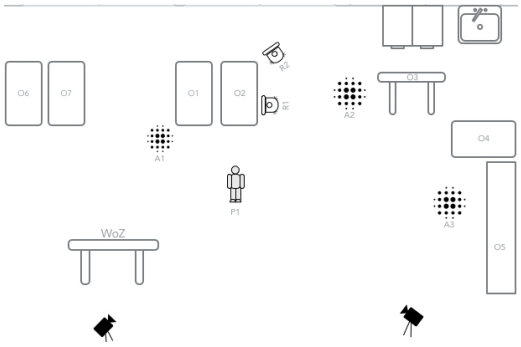

Figure 1: Chemistry laboratory floor schematic

The robot's positioning system, branded '360 LaserSmart Mapping' [37], creates maps of the environment bounded by walls and obstacles however it is not possible to control the navigation system or even access the current position of the robot using presently available APIs. Due to this, the Wizard was constrained by the possible range of responses. The app provided the location of the robot when it was in its base station. The Wizard controlled the robot according to the app's functions, which were 'start cleaning everything', 'clean a pre-programmed zone', 'pause', 'stop', and 'return to base'. In other words, the Wizard could not choose to manually control the robot and the Wizard's actions were constrained to ensure that their "human intelligence [did not] stray too far beyond the performance of the future system" [16]

The study took place over ten days, with each study taking 20–30 minutes. The setting was not 'staged', with equipment positioned as and where we found it, although we ensured this was consistent throughout all studies. Each study was video and audio recorded with two wide-angle cameras on tripod, including a hand-held recorder (to allow for higher quality audio capture).

3.1.2 Description of the Wizard's control. One of the researchers acted as the Wizard during all the studies and was aware of the goals of the study to capture the language and interaction used with the robot. The Wizard did not know the participants prior to the study. Their role was to operate the voice-based software. Throughout the study, the Wizard tried to respond to requests from users as promptly as possible, balancing the act of instructing the robot to complete the task using the iPad app and generating the synthesised response.

The wizard's decision whether to treat a given instruction as 'successful' or not (i.e., to trigger the robot's action desired by the user) was based on several factors: 1) the ability to perform the instruction using the app, 2) in requests to clean a specific area, the ability for the Wizard to precisely understand the request. The lack of live positioning data meant that requests for cleaning relative to the robot's current position and rotation in the environment did not 'succeed' if robot was not in its base station (e.g., 'clean in front of you' is a constantly changing while the robot is moving, but

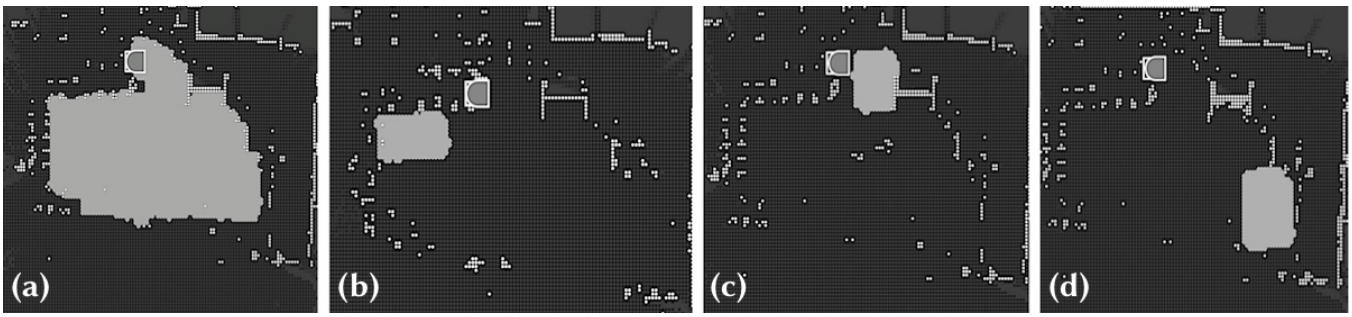

Figure 2: Maps of the environment from the Neato platform (in light grey) of: (a) the entire space, and the three areas to clean: (b) *A1*, (c) *A2*, and (d) *A3*

possible when the robot is in its base in front of a pre-planned area for cleaning). If the Wizard did not understand a specific location, an error response including a partial transcription of the location given by the participant would be generated (e.g., *'Sorry, I don't understand where by the window is'* if the participant asked to clean by the window). If the participant did not provide a specific location, a request for one would be given (e.g., *'Where should I clean?'*).

The Wizard made notes using a piece of paper when a decision was made to respond to a request in a given way, to ensure consistency across the studies and reduce 'Wizarding errors'. As per the intention of the study to elicit the language for instructions, there was not a strict protocol for the Wizard to follow, rather an outline of the scenarios and tasks participants were required to complete. The Wizard responded to requests from participant, recording their decisions, and iteratively evolving the protocol throughout the studies, forming a list of rules that would be applied to subsequent requests [40]. For example, the Wizard accepted requests that did not start with the command 'robovac' despite this being initially planned, and as a result, this change in practice became part of the protocol.

3.1.3 Participants. 21 participants volunteered for the study, 11 self-identifying males and 10 self-identifying females (see Table 1). Participants were contacted through written notices and e-mails within the university's Faculty of Engineering. We recruited students and staff who work in the laboratory at least periodically thus would have familiarity with the environment and equipment found there. After agreeing to taking part in the study, participants completed a demographic questionnaire and answered questions about their cleaning practices and routines in relation to the laboratory facilities.

3.1.4 Procedure. We conducted preliminary interviews and issued a questionnaire to understand participants' prior experience with robots and voice interfaces. Participants then proceeded to interact with the robot through five different pre-developed scenarios that we envisage might come up in such an environment. Participants were asked to instruct the robot as they wish (although we asked them to call/refer to it as *robovac* when issuing instructions). We specifically instructed participants that "there was not a correct or wrong way to do it". As explained above, when the participant gave an instruction to the robot that provided enough information for the Wizard to interpret which area to clean, the robot was commanded by the Wizard to clean the area using the iPad app (see above). After interacting with robot, we conducted an exit interview to understand participants' reflections of their interactions with robot.

Table 1: Approximate age, self-identified gender, and occupation of participants

#	Age	Gender	Occupation
P1	25–29	M	Researcher and Technician
P2	50+	F	Learning Tech. Consultant
P3	35–39	M	Learning Tech. Consultant
P4	50+	F	Learning Tech. Consultant
P5	35–39	F	Chemical and Env. Engineer
P6	50+	F	Laboratory Technician
P7	25–29	F	Laboratory Technician
P8	30–34	F	Laboratory Technician
P9	30–34	M	Laboratory Technician
P10	25–29	F	Student
P11	20–24	M	Student
P12	25–29	M	Building Attendant
P13	35–39	M	Building Attendant
P14	45–49	M	Laboratory Technician
P15	30–34	F	Student
P16	25–29	M	Research
P17	25–29	M	Student
P18	45–49	M	Administrator
P19	30–34	F	Student
P20	25–29	F	Student
P21	25–29	M	Student

3.1.5 Scenarios. Participation was structured and managed in accord with various prepared scenarios. To begin with we will detail a simple example of the basic *clean* scenario, and then describe the other remaining types (*status*, *interrupt*, *queuing* and *schedule*).

Clean The participant is asked to choose one of the three 'piles' of debris available and instruct the robot in a 'natural way' to clean one of those areas. The following vignette provides an example of the kind of dialogue that would unfold.

Vignette. P1 works on research and as a Teaching Technician in the chemistry laboratory (male, 25–29 years

old). They do not own any smart technology at home and their only experience interacting with technology through voice is using an Amazon Alexa. They do not have experience using robots in the workplace. When the researcher asks them to instruct the robot to clean any area in the lab, the following dialogue unfolded.

P1: robovac can you clean the floor over here please?
Robot: sorry, I don't understand where over here is
P1: robovac clean near the table
Robot: ok, I will clean near the table now

Status The participant is asked to request the status of the robovac while it is cleaning (e.g., to find out what the robot is doing, or how long the cleaning task will take).

Interrupt The participant is asked to interrupt the current task of the robot and send it to do another activity (or to return to its starting position).

Queuing The participant is asked to instruct the robovac to do another cleaning task after it has completed the current one (this was introduced to participants as 'queuing up' of tasks).

Schedule The participant is asked to instruct the robovac to clean areas of the laboratory at a future time e.g., after finishing their (i.e., the researcher's) session in the laboratory, or after everyone has left.

3.2 RoboClean dataset description

The data collected from the 21 participants' trials are represented in a single table consisting of the following parameters:

- **scenario** is a numerical index for the scenario the participant is completing and can be any of the integers 1–5, corresponding to the *Clean*, *Status*, *Interrupt*, *Queuing*, and *Schedule* respectively (see above)
- **input message** is the transcriptions of each participants' instructions to robovac
- **input time** provides a chronological order of each instruction
- **output message** is the Wizard's 'response' to each instruction
- **performance** includes the categorisation of successful and unsuccessful instructions followed by the robot (see subsubsection 3.1.2)
- **reference** and **parameter** correspond to the final categorisation of participants' spoken instructions (see section 4)
- **embodied actions** describe all the participants' body movements, labeled following Giuliani et al.'s categories [18]; the dataset includes head and body movements, posture, facial expressions, speech, hand gestures, and emotions

The RoboClean dataset is available under a CC-BY license at https://doi.org/10.17639/nott.7295.

4 UNDERSTANDING SPOKEN INSTRUCTIONS: LINGUISTIC ANALYSIS

This section describes the qualitative linguistic analysis we conducted to identify the categories of instructions and referential parameters that lead to inform the creation of an automated classification process using Machine Learning.

The audio recorded from the robot interactions was transcribed, segmented by scenarios and participants. We collected 525 instructions to the robot and classified the embodied actions from those inputs, based on the video categorisation for social signal suggested by Giuliani et al. [18]. In this, we identified the head movement, body movement/body posture, facial expressions, hand gestures, and speech associated with each participant input. In this paper, we focus on the use of speech only, leaving the other forms of multi-modal interaction as out of scope. Additionally, following a thematic analysis approach [2, 8], two researchers classified how the participants' requests could be parsed by algorithm to determine the desired location the participant wanted cleaning, examining the referential categories used. In other words, our goal was to categorise how the participant specified the location of the position to clean. For this process, both researchers independently classified each request to zero or more positional references. After that, discrepancies were reviewed and agreement sought. Below we detail the classification of these 525 inputs into four referential categories. Following the referential categories, one researcher grouped data into 5 different parameters related to the scope of each referential category.

4.1 Spoken instructions in context

Here we present results from our analysis of participants' spoken instructions *in situ*. In general, the grammatical logic of an instruction in the English language follows the pattern **Address + Action + Parameter**. The *Address* term (e.g., "robovac", or "vacuum robot") (which may be omitted), the *Action* term corresponds to the predicate (verb) (sometimes formulated as a question), and the *parameter* corresponds to the object of the instruction. To understand what kind of instructions a voice-based natural language interface would need to support in future, in what follows we focus on how people construct the object / parameter part of the instruction.

4.1.1 Referential categorisation of instructions. As instructions are constructed in a *situated* manner in reference to the setting and tasks, it matters to examine them in relation to how the instructions were constructed by participants. Our thematic analysis yielded five top-level referential categories:

- **Space**: relative to the environment in some way *e.g "robovac clean this whole area", "robovac clean the table area and the camera area"*
- **Operator**: relative to the person giving the commands *e.g "robovac come and clean next to where I am"*
- **Base**: relative to the base *e.g "robovac can you go home"*
- **Robot**: using the robot's current positioning to go somewhere, rather than the robot being told to go to something based on a point in the environment *e.g "robovac stop what you're doing and clean directly behind you"*
- **Other**: no positioning, uncertain information. *e.g "robovac start", "robovac what time is it now?"*

Most of the instructions were constructed relative to the robot itself (32.38%) and the space (21.90%). It was less common that instructions were constructed relative to the operator themselves (1.71%) and the base (2.10%). The category "other" captures the remaining 35.62% of instructions and it includes instructions related

to questions to the robot, greetings and verbs without any additional information of the environment. Additionally, there were combinations of the main categories, e.g., robot-operator, robot-space, space-robot and so on, which in total corresponds to 6.29% of categories.

4.1.2 Parameter construction. Based on further thematic analysis of participants' instructions, we examined in more detail the grammatical objects used in the referential instructions to "parameterise" the instruction, inspired by Wobbrock et al. [61]'s gesture taxonomy. Table 2 shows the parameters and frequencies distributed by scenarios. We included just the four main referential categories (305 instructions) excluding the category "other" (187 instructions) and combination of "space: operator, robot, base", "robot: operator, space, base" and "operator: robot, space" (36 instructions in total). We excluded from the analysis the instructions categorized as "other" because these instructions did not contain referential information. Some examples of the category *other* were for example *robovac start; robovac can you clean for me?; robovac next; stop; robovac how dirty is it?*. Most of the instructions in the category "other" were related to the "Status" scenario, reflecting that this scenario was generating bidirectional communication or encouraging a sort of dialogue with the robot instead of human-to-robot unidirectional instructions.

Table 2 shows that *object-centric* instructions are mostly related to the *space* rather than *robot*. This reflects that when the instruction is constructed in reference to the robot's current position it does not contain more details about the elements in the environment. We identified a difference as well between *object-centric* and *room-centric*, with participants delegating cleaning tasks using as references particular objects inside the room that define an area to be cleaned (e.g., *clean by the water bath, clean under the tank, clean next to the bin*), but without specifying how much surface should be covered. In contrast, *room-centric* comprises typically more coarse-grained instructions for cleaning larger areas e.g., *clean the whole lab, clean the whole room, clean everywhere.*

Table 2: Contextual parameters of instructions for each scenario and in total

Reference	Parameter	S1	S2	S3	S4	S5	∑
Robot	Direction	52	15	55	34	1	157
	Distance, Direction	1		2		2	5
	Distance		1	3			4
	Direction, Time				1	2	3
	Direction, Room-Centric			1			1
Space	Object-Centric	7	6	16	14	6	49
	Room-Centric	4	2	5	5	21	37
	Room-Centric, Time	1				18	19
	Object-Centric, Time				1	6	7
	Direction, Time					1	1
	Room-Centric, Object-Centric		1				1
	Object-Centric, Distance, Direction				1		1
Operator	Direction	3	1	2	1		7
	Distance	1					1
	Direction, Time					1	1
Base	Direction	1		9		1	11

Figure 3 presents the most common *instruction types* that emerge from the data with examples for each one. We obtain instruction types when we combine reference and parameter, i.e., Table 2 lists 16 types. The most frequent instruction type was *Robot-Direction* (157/170), which was typically found in the "Clean" and "Interrupt" scenarios.

For instructions constructed in reference to *Space*, the most frequent parameter was *object-centric* to reference nearby objects, e.g., *by tanks, by the pipes, behind this desk, near to the window.*

5 DISCUSSION AND DESIGN IMPLICATIONS

Following the collection of participants' instructions to robots through our five scenarios, we identified the core linguistic elements of instructions to the robot. We then categorised the different ways in which people formulated these instructions *in situ*, by drawing out the different referential categories and various contextual parameters participants used in our study. This contextual information is a key source of data, providing the robot with additional information to 'understanding' both the environment and the instructions given to it by operators, to improve information exchange [39] and effective communication [29, 57]. This serves to support richer dialogues that may contribute to more fluent natural language interactions [9] and maximising verbal understanding [30, 57], specially in constrained environments when training data might be limited for language-based human-robot interactions [38, 56].

In our work, we found the terminology people tended to use in the instructions was (of course) constructed in a contextual manner that took the present situation into account. More specifically than this, we identified a way to break down this contextualisation into something much more tractable for adoption in the development of natural language interfaces for robots in specialised environments. To this end we offer a grounded schema for dealing with referential categories, in particular the following ways of cutting up the space of language: references to the environment, references to the position of the robot, its base, and the operator themselves.

We now elaborate on these in terms of their *robot-centricity*, *object-centricity* and *room-centricity*.

5.1 Robot-centric instructions

For navigating the world, mobile robots commonly use referring information [57], establishing a link between robot's information position and what is included in the referential language of the instruction in addition to internal planning and navigation algorithms. Robot-centric instructions focused on direction and distance but seemed to take for granted the robot's location. In our data, the utterances related to robot position represent 32.38% of the instructions, and of that, more than 90% includes direction as a contextual parameter. The instructions reflect that having robot-centric spoken instructions will require a robot to know and use its location. Hence, the starting point for any robot-centric instruction will rely on robot understanding of spatial relationships (position). For example, the instruction "robovac can you clean five feet to your right" builds on the assumption of environmental understanding and previous knowledge of where the robot is physically located in space. The robot should recognise their front, back, left, right, up and down before combining it with any additional source of data that complements their own "robot self positional awareness" to act. The spatial language identification at this level of detail would

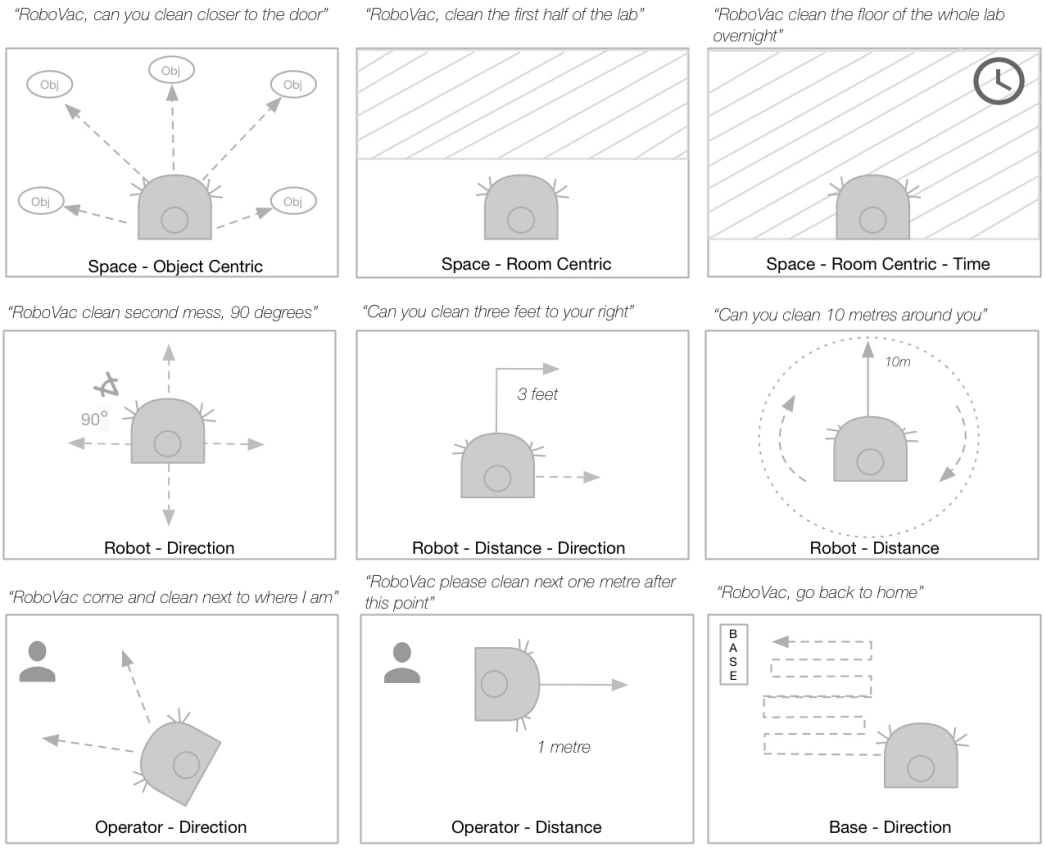

Figure 3: Most common instruction types: combinations of referential categories with contextual parameters

contribute to improving the robot communicative understanding process [29], providing mechanisms to identify the instructions received and to prioritise its position to then combine with contextual parameters.

5.2 Object-centric instructions

The *contextual information* related to the environment was mainly expressed in terms of objects and areas within the room. 21.9% of the instructions were related to the space, of which 42.6% correspond to "object-centric" instructions. For example, the space-object type instruction "robovac can you clean by the grey frame over there please" articulates a target area in which to clean by drawing on an environmental reference in terms of an object (the grey frame). The participants' local knowledge especially came to the fore regarding the object descriptions. This contextual information is crucial to be available to process contextually-relevant terms about objects within a given space, many of which have multiple identifiers such as variants of product names and types. In our experiment—in the chemistry laboratory—examples of such objects include "bleed gauge", "valve board", "grey tanks", "water baths", and "the rig". This kind of variable local knowledge and terminology used to reference objects in the present environment will need to be present in future systems so as to support greater immediate familiarity with robotics. This implies the need for SLU to be configurable with the contextual terminology about objects that are part of those spaces, their names and ways of referencing them that are reflective of local working cultures and practices: e.g., "grey tanks", "water baths" and "tank" are 'the same' for our setting.

5.3 Room-centric instructions

In contrast with the previous two types, for room-centric instruction types (32.17% of instructions related to space), participants made use of common terminology to refer to specific areas of the laboratory. For example, "robovac clean the whole room" delineates areas of the surface to clean by using terms such as "whole", "all", "everywhere", and "half of the lab". This raises particular challenges; while people who use such spaces understand which area of the room is considered "half of the lab", where this "half" starts and where it ends poses a difficult challenge to disambiguate and do so appropriately for the given situation.

The referential categories most used by participants included direction, distance, and time e.g., "clean right" or "turn 180 degrees", and so on. In the case of distance, users' instructions typically included a particular point to go to, e.g., "travel one metre and 10 metres diagonally". This reinforces the challenge in implementing a system that can handle specifics where given but also where they are omitted.

5.4 Challenges for contextual understanding in low-resource environments

There are several significant challenges to overcome to provide a rich source of knowledge that helps to bridge the gap from "voice to action" for the effective instruction of robots [29, 50]. In our case study, one of the challenges that emerges is how to address the contextual information in specialised low-resource environments.

To advance HRI from more simple pre-defined commands to dynamic/flexible interactions, it is critical also to understand the context of the interactions [56], such as where does the language of required actions come from and how does it relate to the situated robot [30, 38, 50, 57]. As Mavridis [30] highlights, it is vital to integrate "language with sensing, action, and purpose in conversational robots". However, to achieve this it is necessary to connect the language with the physical context, known as "situated language". Such situated SLU would need to be built in a systematic, situationally-specific way, thus providing a starting point for makers of, for instance, future situated language models in other kinds of specific environments. We offer some preliminary thoughts how this might be achieved in the following.

5.5 Automating situated language understanding

Finally, we reflect on future work needed to extend our work towards situated SLU for interacting with voice-controlled robots.

Providing a common ground of understanding is critical for improving the interaction between humans and robots [30, 50, 57]. In our study, we were able to identify the referential categories and parameters of spoken instructions.

The next steps would include separating the instructions into multi-label and binary problems to evaluate the feasibility of automatic labelling based on performance on different models. This learning process enables the generation of a metric to evaluate the linguistic model, allowing the protocol to be iterated until a desired learning accuracy threshold is reached. Following this, integration in the speech-based interactive design of a robot can be completed.

In this paper we have outlined the first steps of what might be termed a 'protocol' for natural language grounding and suggest that this protocol could be applied in different specialised environments beyond the specific setting we investigated. The various stages of this protocol would enable the generation and modelling of situationally-relevant language for a range of different service robots in diverse specialised environments—although we must caveat this to say that future work is needed to determine the precise contours of how generalisable this protocol actually is.

First, we want outline what the key stages of this protocol might be, in order to provide a clear 'toolkit' for others:

(1) **Task-Setting**. Defining the task(s) to be performed by the robot(s). Through the definition of these tasks, situational aspects also begin to be crafted, e.g., that tasks *A*, *B*, and *C* are usually undertaken in the particular setting.
(2) **Bootstrapping**. For situations in which there is no extant situationally-specific language data for the given setting, we recommend a bootstrapping process, consisting of:

    Elicitation. Gathering the language employed in situationally-specific tasks through the design and orchestration of simple scenarios where potential operators who possess local knowledge—such as specific terminology for the objects and spaces in the setting—provide spoken instructions to the robot, such as through a Wizard of Oz (WOz) study similar to the one we demonstrated in our paper.

    Data Augmentation. To make the reasonably sparse dataset thus collected more robust for modelling, carry out a data augmentation process. Augmenting our dataset through permutation of words allows our classifier to be less sensitive to those variations in language.
(3) **Linguistic analysis**. Carry out a linguistic analysis to understand the grammatical composition, terminology, and potential categories in the elicited natural language instructions. The outputs of this serve as labels for the modelling step.
(4) **Modelling**. Generating Machine Learning models where metrics can serve to evaluate the potential applicability of the model in the setting. Repeat steps (1)-(3) until the desired level of accuracy has been achieved.

One of the advantages of the protocol may be to apply it where data collection can be logistically challenging, e.g., because of cost or time constraints, non English-centric environments, as well as necessitating setting-specific vocabulary. In particular, this approach to generating Machine Learning models for low-resource environments could be accurate with a relatively *small* quantity of data (as opposed to the relatively 'big' datasets normally required to generate such models).

5.6 Limitations and Future Work

Our study does have some limitations. First, this is a partial contribution to the grounded learning problem, our study focuses only on the semantic representation of instructions, without including objects or image representation and recognition. Second, there are remaining challenges that we did not explore such as integrating *ambiguity* and *unstructured sentences* of the instructions within the grounded schema for specialised low-resource environment. Third, the methodology used in this study is not new; however, applying a qualitative approach to elicitate specific terminology makes it possible to contribute a contextualised dataset for HRI in data-constrained environments (low-resource contexts). Finally, future work can include different data constrained contexts, allowing WOz removal, and recognising patterns of successful spoken instructions without a Wizard, focusing on details of successful instructions such as speech tone, speed and intonation.

6 CONCLUSION

As robots controlled through voice are becoming more pervasive it is important to understand how to design for situationally-appropriate spoken language understanding in all kinds of environments. This work sought to provide an understanding of situational language use in robot instructions in a specialised environment; and we discuss how the resulting dataset could be used to contribute to automate situated language understanding.

Our work in this paper contributes a grounded language dataset in a low-resource environment. HRI designers and researchers can use this dataset as a starting point to build on and refine to create their own situated natural language interfaces that are bespoke to any specific environment, by example of the RoboClean case study involving instructions of a robot. Through a WOz study we elicited the specialised language *in situ* through which participants with local expert knowledge, such as terminology for what objects and areas in the space are called, instructed a voice-controlled robot. We suggest that this approach is, in principle, applicable in other specialised environments and should enable those wishing to build natural language robot interaction for those environments.

Our qualitative analysis of more than 500 spoken instructions of the grammar of instructions (address, action and parameter) suggests that it is important that specific local terminology can be understood for instructions to be successful, supporting the case for future work in this area. We then discuss a number of additional tasks that would be needed to explore whether and how spoken language understanding of instructions in specialised environments could be automated, which would need to be validated in future work, particularly with non-WOz voice-controlled robots.

ACKNOWLEDGMENTS

This work was supported by the Engineering and Physical Sciences Research Council [grant numbers EP/M02315X/1 (https://gtr.ukri.org/projects?ref=EP%2FM02315X%2F1), EP/R045127/1 (https://gtr.ukri.org/projects?ref=EP%2FR045127%2F1), EP/N014243/1 (https://gtr.ukri.org/projects?ref=EP%2FN014243%2F1), EP/V00784X/1 (https://gtr.ukri.org/projects?ref=EP%2FV00784X%2F1)]. We would like to thank Charlotte Gray and Jane Slinger for their help in collecting and labeling data.

REFERENCES

[1] Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz. 2014. Automatic speech recognition for under-resourced languages: A survey. *Speech communication* 56 (2014), 85–100.

[2] Ann Blandford, Dominic Furniss, and Stephann Makri. 2016. Qualitative HCI Research: Going Behind the Scenes. *Synthesis Lectures on Human-Centered Informatics* 9, 1 (2016), 1–115. https://doi.org/10.2200/S00706ED1V01Y201602HCI034

[3] Claire Bonial, Matthew Marge, Ron artstein, Ashley Foots, Felix Gervits, Cory J. Hayes, Cassidy Henry, Susan G. Hill, Anton Leuski, Stephanie M. Lukin, Pooja Moolchandani, Kimberly A. Pollard, David Traum, and Clare R. Voss. 2017. Laying Down the Yellow Brick Road: Development of a Wizard-of-Oz Interface for Collecting Human-Robot Dialogue. arXiv:1710.06406 [cs.CL]

[4] BRCGS. 2018. BRC Global Standard for Food Safety F804a: Issue 8 - Auditor Checklist and Site Self-Assessment Tool. https://www.brcglobalstandards.com/media/1055378/food-safety-issue-8-checklist-english.docx

[5] Jian-Hua Cai. 2017. Near-Infrared Spectrum Detection of Wheat Gluten Protein Content Based on a Combined Filtering Method. *Journal of AOAC International* 100, 5 (2017), 1565–1568. https://doi.org/10.5740/jaoacint.17-0008

[6] Leigh Clark, Philip Doyle, Diego Garaialde, Emer Gilmartin, Stephan Schlögl, Jens Edlund, Matthew Aylett, João Cabral, Cosmin Munteanu, Justin Edwards, and Benjamin R Cowan. 2019. The State of Speech in HCI: Trends, Themes and Challenges. *Interacting with Computers* 31, 4 (09 2019), 349–371. https://doi.org/10.1093/iwc/iwz016 arXiv:https://academic.oup.com/iwc/article-pdf/31/4/349/32471570/iwz016.pdf

[7] Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, Vincent Wade, and Benjamin R Cowan. 2019. What Makes a Good Conversation? Challenges in Designing Truly Conversational Agents. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. ACM, New York, NY, USA, Article 475, 12 pages. https://doi.org/10.1145/3290605.3300705

[8] John W Creswell and Vicki L Plano Clark. 2017. *Designing and Conducting Mixed Methods Research*. SAGE Publications, Los Angeles, CA, USA. 520 pages.

[9] Heriberto Cuayáhuitl. 2015. Robot learning from verbal interaction: A brief survey. In *AISB Convention 2015*. Society for the Study of Artificial Intelligence and the Simulation of Behaviour, London, 4.

[10] Fethiye Irmak Doğan and Iolanda Leite. 2021. Open Challenges on Generating Referring Expressions for Human-Robot Interaction. *arXiv preprint arXiv:2104.09193* (2021).

[11] W Randolph Ford and Raoul N Smith. 1982. Collocational Grammar As a Model for Human-computer Interaction. In *Proceedings of the 9th Conference on Computational Linguistics - Volume 2 (COLING '82)*. Academia Praha, Czechoslovakia, 106–110. https://doi.org/10.3115/990100.990122

[12] Jodi Forlizzi. 2007. How Robotic Products Become Social Products: An Ethnographic Study of Cleaning in the Home. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction* (Arlington, Virginia, USA) *(HRI '07)*. ACM, New York, NY, USA, 129–136. https://doi.org/10.1145/1228716.1228734

[13] Jodi Forlizzi and Carl DiSalvo. 2006. Service Robots in the Domestic Environment: A Study of the Roomba Vacuum in the Home. In *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction* (Salt Lake City, Utah, USA) *(HRI '06)*. ACM, New York, NY, USA, 258–265. https://doi.org/10.1145/1121241.1121286

[14] Fortune Business Insights. 2019. Robotic Vacuum Cleaners Market Size, Share and Industry Analysis By Type. Retrieved April 7, 2020 from https://www.fortunebusinessinsights.com/industry-reports/robotic-vacuum-cleaners-market-100645

[15] Mary Ellen Foster. 2019. Natural language generation for social robotics: opportunities and challenges. *Philosophical Transactions of the Royal Society B* 374, 1771 (2019), 20180027.

[16] Norman M Fraser and Nigel Gilbert. 1991. Simulating speech systems. *Computer Speech & Language* 5, 1 (jan 1991), 81–99. https://doi.org/10.1016/0885-2308(91)90019-M

[17] Helena Anna Frijns, Oliver Schürer, and Sabine Theresia Koeszegi. 2021. Communication Models in Human–Robot Interaction: An Asymmetric MODel of ALterity in Human–Robot Interaction (AMODAL-HRI). *International Journal of Social Robotics* (2021). https://doi.org/10.1007/s12369-021-00785-7

[18] Manuel Giuliani, Nicole Mirnig, Gerald Stollnberger, Susanne Stadler, Roland Buchner, and Manfred Tscheligi. 2015. Systematic analysis of video data from different human–robot interaction studies: a categorization of social signals during error situations. *Frontiers in Psychology* 6, Article 931 (2015), 12 pages. https://doi.org/10.3389/fpsyg.2015.00931

[19] Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena* 42, 1-3 (1990), 335–346.

[20] Tero Jokela, Parisa Pour Rezaei, and Kaisa Väänänen. 2016. Using Elicitation Studies to Generate Collocated Interaction Methods. In *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct* (Florence, Italy) *(MobileHCI '16)*. ACM, New York, NY, USA, 1129–1133. https://doi.org/10.1145/2957265.2962654

[21] Malte Jung and Pamela Hinds. 2018. Robots in the Wild: A Time for More Robust Theories of Human-Robot Interaction. *J. Hum.-Robot Interact.* 7, 1, Article 2 (May 2018), 5 pages. https://doi.org/10.1145/3208975

[22] Gunhee Kim, Woojin Chung, Kyung Rock Kim, Munsang Kim, Sangmok Han, and Richard H. Shinn. 2004. The autonomous tour-guide robot Jinny. In *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Vol. 4. IEEE, Piscataway, NJ, USA, 3450–3455. https://doi.org/10.1109/IROS.2004.1389950

[23] Thomas Kollar, Stefanie Tellex, Deb Roy, and Nicholas Roy. 2010. Toward understanding natural language directions. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, Piscataway, NJ, USA, 259–266. https://doi.org/10.1109/HRI.2010.5453186

[24] Maret Kullasaar, Evely Vutt, and Mare Koit. 2002. Developing a natural language dialogue system: Wizard of Oz studies. In *Proceedings First International IEEE Symposium Intelligent Systems*, Vol. 1. IEEE, New York, NY, USA, 202–207. https://doi.org/10.1109/IS.2002.1044255

[25] Nicolas Lair, Clement Delgrange, David Mugisha, Jean-Michel Dussoux, Pierre-Yves Oudeyer, and Peter Ford Dominey. 2020. User-in-the-Loop Adaptive Intent Detection for Instructable Digital Assistant. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Cagliari, Italy) *(IUI '20)*. Association for Computing Machinery, New York, NY, USA, 116–127. https://doi.org/10.1145/3377325.3377490

[26] Gregory Lemasurier, Gal Bejerano, Victoria Albanese, Jenna Parrillo, Holly A. Yanco, Nicholas Amerson, Rebecca Hetrick, and Elizabeth Phillips. 2021. Methods for Expressing Robot Intent for Human–Robot Collaboration in Shared Workspaces. *J. Hum.-Robot Interact.* 10, 4, Article 40 (Sept. 2021), 27 pages. https://doi.org/10.1145/3472223

[27] Luc. 2012. *Grounding Language through Evolutionary Language Games*. Springer US, Chapter 1, 1–20.

[28] James Manyika. 2017. *A future that works: AI, automation, employment, and productivity*. Technical Report. McKinsey Global Institute Research.

[29] Matthew Marge, Carol Espy-Wilson, Nigel G. Ward, Abeer Alwan, Yoav Artzi, Mohit Bansal, Gil Blankenship, Joyce Chai, Hal Daumé, Debadeepta Dey, Mary Harper, Thomas Howard, Casey Kennington, Ivana Kruijff-Korbayová, Dinesh

Manocha, Cynthia Matuszek, Ross Mead, Raymond Mooney, Roger K. Moore, Mari Ostendorf, Heather Pon-Barry, Alexander I. Rudnicky, Matthias Scheutz, Robert St. Amant, Tong Sun, Stefanie Tellex, David Traum, and Zhou Yu. 2022. Spoken language interaction with robots: Recommendations for future research. *Computer Speech & Language* 71 (2022), 13. https://doi.org/10.1016/j.csl.2021.101255

[30] Nikolaos Mavridis. 2015. A review of verbal and non-verbal human–robot interactive communication. *Robotics and Autonomous Systems* 63 (2015), 22–35. https://doi.org/10.1016/j.robot.2014.09.031

[31] Michael McTear, Zoraida Callejas, and David Griol. 2016. *The Conversational Interface: Talking to Smart Devices.* Springer International Publishing, Cham, Switzerland. 422 pages.

[32] Michael F McTear. 2002. Spoken dialogue technology: enabling the conversational user interface. *ACM Computing Surveys (CSUR)* 34, 1 (2002), 90–169.

[33] Josh Meyer. 2019. *Multi-task and transfer learning in low-resource speech recognition.* Ph.D. Dissertation. The University of Arizona.

[34] Yasser Mohammad. 2018. Natural Human-Robot Interaction. *The Wiley Handbook of Human Computer Interaction* 2 (2018), 641–655.

[35] Shiwali Mohan. 2015. *From Verbs to Tasks: An Integrated Account of Learning Tasks from Situated Interactive Instruction.* Ph.D. Dissertation. The University of Michigan.

[36] Neato Robotics. 2019. Neato Botvac D7. Retrieved April 8, 2020 from https://www.neatorobotics.com/gb/robot-vacuum/d-shape-series/neato-d7/

[37] Neato Robotics. 2019. Neato Botvac D7 LaserSmart. Retrieved April 8, 2020 from https://www.neatorobotics.com/gb/lasersmart-mapping-navigation/

[38] José Novoa, Rodrigo Mahu, Jorge Wuth, Juan Pablo Escudero, Josué Fredes, and Néstor Becerra Yoma. 2021. Automatic Speech Recognition for Indoor HRI Scenarios. *J. Hum.-Robot Interact.* 10, 2, Article 17 (March 2021), 30 pages. https://doi.org/10.1145/3442629

[39] José Novoa, Jorge Wuth, Juan Pablo Escudero, Josué Fredes, Rodrigo Mahu, and Néstor Becerra Yoma. 2018. DNN-HMM Based Automatic Speech Recognition for HRI Scenarios. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (Chicago, IL, USA) *(HRI '18)*. ACM, New York, NY, USA, 150–159. https://doi.org/10.1145/3171221.3171280

[40] Martin Porcheron, Joel E Fischer, and Stuart Reeves. 2021. Pulling Back the Curtain on the Wizards of Oz. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW3, Article 243 (jan 2021), 22 pages. https://doi.org/10.1145/3432942

[41] Martin Porcheron, Joel E Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice Interfaces in Everyday Life. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. ACM, New York, NY, USA, Article 640, 12 pages. https://doi.org/10.1145/3173574.3174214

[42] Martin Porcheron, Joel E Fischer, and Michel Valstar. 2020. NottReal: A tool for voice-based Wizard of Oz studies. In *Proceedings of the 2nd Conference on Conversational User Interfaces* (Virtual Event) *(CUI '20)*. ACM, New York, NY, USA, Article 35, 3 pages. https://doi.org/10.1145/3405755.3406168

[43] Chloé Pou-Prom, Stefania Raimondo, and Frank Rudzicz. 2020. A conversational robot for older adults with alzheimer's Disease. *ACM Transactions on Human-Robot Interaction (THRI)* 9, 3 (2020), 1–25.

[44] Abhinanda R Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J Black. 2021. BABEL: Bodies, Action and Behavior with English Labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 722–731.

[45] L Rabiner and S Levinson. 1985. A speaker-independent, syntax-directed, connected word recognition system based on hidden Markov models and level building. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 33, 3 (1985), 561–573.

[46] Ahmed Rady, Joel Fischer, Stuart Reeves, Brian Logan, and Nicholas James Watson. 2020. The Effect of Light Intensity, Sensor Height, and Spectral Pre-Processing Methods When Using NIR Spectroscopy to Identify Different Allergen-Containing Powdered Foods. *Sensors* 20, 1 (2020). https://doi.org/10.3390/s20010230

[47] Stuart Reeves. 2019. Conversation Considered Harmful?. In *Proceedings of the 1st International Conference on Conversational User Interfaces* (Dublin, Ireland) *(CUI '19)*. Association for Computing Machinery, New York, NY, USA, Article 10, 3 pages. https://doi.org/10.1145/3342775.3342796

[48] Laurel D Riek. 2012. Wizard of Oz Studies in HRI: A Systematic Review and New Reporting Guidelines. *J. Hum.-Robot Interact.* 1, 1 (July 2012), 119–136. https://doi.org/10.5898/JHRI.1.1.Riek

[49] Robotic Industries Association. 2020. Industrial Cleaning Robots. Retrieved April 7, 2020 from https://www.robotics.org/service-robots/industrial-cleaning-robots

[50] Raquel Ros, Séverin Lemaignan, E Akin Sisbot, Rachid Alami, Jasmin Steinwender, Katharina Hamann, and Felix Warneken. 2010. Which one? grounding the referent based on efficient human-robot interaction. In *19th International Symposium in Robot and Human Interactive Communication*. IEEE, Piscataway, NJ, USA, 570–575. https://doi.org/10.1109/ROMAN.2010.5598719

[51] Allison Sauppé and Bilge Mutlu. 2015. The Social Impact of a Robot Co-Worker in Industrial Settings. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. ACM, New York, NY, USA, 3613–3622. https://doi.org/10.1145/2702123.2702181

[52] Monica Schofield. 1999. "Neither master nor slave...". A practical case study in the development and employment of cleaning robots. In *1999 7th IEEE International Conference on Emerging Technologies and Factory Automation. Proceedings ETFA '99 (Cat. No.99TH8467)*, Vol. 2. IEEE, Piscataway, NJ, USA, 1427–1434. https://doi.org/10.1109/ETFA.1999.813157

[53] Jan Smeddinck, Kamila Wajda, Adeel Naveed, Leen Touma, Yuting Chen, Muhammad Abu Hasan, Muhammad Waqas Latif, and Robert Porzel. 2010. QuickWoZ: A Multi-Purpose Wizard-of-Oz Framework for Experiments with Embodied Conversational Agents. In *Proceedings of the 15th International Conference on Intelligent User Interfaces* (Hong Kong, China) *(IUI '10)*. Association for Computing Machinery, New York, NY, USA, 427–428. https://doi.org/10.1145/1719970.1720055

[54] Dimitris Spiliotopoulos, Ion Androutsopoulos, and Constantine D Spyropoulos. 2001. Human-robot interaction based on spoken natural language dialogue. In *Proceedings of the European Workshop on Service and Humanoid Robots*. University of Athens, Athens, Greece, 5.

[55] Arjun Srinivasan, Mira Dontcheva, Eytan Adar, and Seth Walker. 2019. Discovering Natural Language Commands in Multimodal Interfaces. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Marina del Ray, California) *(IUI '19)*. Association for Computing Machinery, New York, NY, USA, 661–672. https://doi.org/10.1145/3301275.3302292

[56] Stefanie Tellex, Nakul Gopalan, Hadas Kress-Gazit, and Cynthia Matuszek. 2020. Robots that use language. *Annual Review of Control, Robotics, and Autonomous Systems* 3 (2020), 25–55.

[57] Stefanie Tellex, Thomas Kollar, Steven Dickerson, Matthew R Walter, Ashis Gopal Banerjee, Seth Teller, and Nicholas Roy. 2011. Understanding Natural Language Commands for Robotic Navigation and Mobile Manipulation. In *Twenty-Fifth AAAI Conference on Artificial Intelligence*. AAAI, Menlo Park, CA, USA, 1507–1514.

[58] Florian Vaussard, Julia Fink, Valérie Bauwens, Philippe Rétornaz, David Hamel, Pierre Dillenbourg, and Francesco Mondada. 2014. Lessons learned from robotic vacuum cleaners entering the home ecosystem. In *Robotics and Autonomous Systems*. Elsevier, Amsterdam, Netherlands, 376–391. https://doi.org/10.1016/j.robot.2013.09.014

[59] Maria Enrica Virgillito. 2017. Rise of the Robots: Technology and the Threat of a Jobless Future. *Labor History* 58, 2 (2017), 240–242.

[60] Christopher David Wallbridge, Séverin Lemaignan, Emmanuel Senft, and Tony Belpaeme. 2019. Generating Spatial Referring Expressions in a Social Robot: Dynamic vs Non-Ambiguous. *Frontiers in Robotics and AI* 6 (2019), 67. https://doi.org/10.3389/frobt.2019.00067

[61] Jacob O Wobbrock, Meredith Ringel Morris, and Andrew D Wilson. 2009. User-Defined Gestures for Surface Computing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, MA, USA) *(CHI '09)*. ACM, New York, NY, USA, 1083–1092. https://doi.org/10.1145/1518701.1518866

[62] Fei Wu et al. 2020. *Child Speech Recognition as Low Resource Automatic Speech Recognition.* Ph.D. Dissertation. Johns Hopkins University.