A Corpus-based Analysis of Route Instructions in Human-Robot Interaction

Theodora Koulouri and Stanislao Lauria

***Abstract*—This paper investigates how users employ spatial descriptions to navigate a speech-enabled robot. We created a simulated environment in which users gave route instructions in a dialogic real-time interaction with a robot, which was operated by naïve participants. The ability of robot monitoring was also manipulated in two experimental conditions. The results provide evidence that the content of the instructions and strategies of the users vary depending on the conditions and demands of the interaction. As expected, the route instructions frequently were underspecified and arbitrary. The findings of this study elucidate the complexity in interpreting spatial language in HRI. However, they also point to the need for endowing mobile robots with richer dialogue resources to compensate for the uncertainties arising from language as well as the environment.**

I. Introduction

A. Natural Language in Human-Robot Interaction

Natural language is arguably the most intuitive and expressive means of communication, stimulating research towards endowing artificial agents with Natural Language Interfaces (NLIs) [1]. In the area of spoken dialogue systems, many commercially successful systems have been developed, which are typically information-retrieving applications based on the slot-filling paradigm [2]. However, Human-Robot Interaction (HRI) is embodied interaction, in which humans and robots coordinate their physical and verbal actions sharing time and space. Evidently, situated dialogue entails greater complexity.

The inherent creativity and ambiguity of natural language along with the poor performance of speech recognition technologies is a general problem in the development of NLIs. Moreover, people are generally uninformed of what robots can do or understand, leading to requests beyond their functional and linguistic domain [3], [4]. Physical co-presence is also expected to reinforce people's perception of common ground and shared knowledge increasing the use of elliptical and underspecified language [5]. In addition, due to requirements of computing power, the capabilities of the robots are also decided upon trade-offs; for instance, endowing a robot with features such as mobility and vision would probably bring restraints to its linguistic abilities [6].

T. Koulouri is with the Department of Information Systems and Computing, Brunel University, Uxbridge, Middlesex, UB8 3PH UK (e-mail: theodora.koulouri@brunel.ac.uk).

S. Lauria, is with the Department of Information Systems and Computing, Brunel University, Uxbridge, Middlesex, UB8 3PH UK (e-mail: stasha.lauria@brunel.ac.uk).

B. Dialogue-based Navigation

A growing and exciting arena of applications is speech-enabled navigating robots. The characteristics of spatial language have been extensively investigated, typically in monologic settings, in the fields of human communication and cognition as well as HRI. Route instructions have an intrinsically linear structure; they typically start with a description of the initial position, then present a segmented series of actions and finally an orientation towards the destination [7]. They generally consist of action descriptions and references to landmarks. Action descriptions include spatial elements, such as direction of motion. References to landmarks are usually supplemented with disambiguating expressions such as "the *second* building *on the left*". Thus, the robot is required to parse and interpret the linguistic content of the instructions, model the actions and descriptions and finally reproduce these actions in the world [8]. However, route instructions by humans are structurally underspecified and arbitrary and involve the application of multiple layers of discourse and situational context. Too often a landmark reference is ambiguous, does not match the internal map of the robot or is out of its vocabulary. Instructions can also lack a clear or systematic point of reference and other important components such as termination points. In addition, there is great individual variation in terms of granularity and details [9], [10]. It has also been observed that users tend to produce vague final instructions that assume a human-like vision in robots (e.g., "turn left and you'll see your destination.") [3]. Therefore, dialogue-based navigation is a highly challenging enterprise for robots that entails understanding of language, spatial actions and relations as well as perception of the world. Incrementally issuing simple commands (for instance, "go straight, and now stop") could moderate the problem but it also compromises the efficiency and naturalness of the interaction. But most significantly, it presupposes constant monitoring and a degree of omniscience by the user which is not possible for all applications.

C. Shared Visual Space in HRI

Numerous studies have investigated the influence of shared visual space in task-oriented human interaction [11]-[13]. In particular, sharing visual information offers direct observation of task status, the addressee's understanding and actions as well as joint focus of attention and reference. This leads to more efficient interactions compared to speech-only settings, and, most importantly, shapes the communication

patterns of the interlocutors. Nevertheless, the role of other co-occurring factors, such as working side-by-side, eye contact, facial expressions and hand gestures, has not been clarified [14]. It also remains an open question whether and in what ways visual information influences the conversational strategies of the user in HRI. As mentioned above, this has several implications for robots that operate with partial or no supervision by the users in collaborative tasks.

D. *Aim of the Study*

The broad aim of the study is to develop a natural framework of communication between a human and a speech-enabled mobile robot. The platform and test-bed for the study is a personal robot which is able to perform and learn navigation tasks by means of unconstrained natural language. This robot is based on the Instruction-Based Learning (IBL) project [15]. Following an empirical approach, our work assumes two perspectives and explores the dialogue behaviour of both partners in the interaction. In particular, previous work by the authors focused on feedback and repair strategies for the robot [16], [17]. Instead, the present paper investigates the linguistic resources employed by users when they collaborate with the robot in a navigation task, with particular interest in their spatial descriptions. Moreover, inspired by findings in human communication (Section C), it also aims to identify the differences in the users' strategies when they can or cannot monitor the robot.

II. METHOD

A. *Experimental Design*

Motivated by previous studies in route instructions [10], [18], [19], we performed a Wizard of Oz study. This approach could help us obtain information on the range of utterances that users spontaneously produce when interacting with a robot and also specify task and system requirements. In an effort to minimise experimental bias and collect as naturally-occurring dialogues as possible, the "wizards" were also naive participants. The domain of the task was navigation and the user had to guide the wizard to several destinations in a simulated town. As explained above, the study also explores the effect of visual information and monitoring on the linguistic choices of the users. Thus, the experimental design involved two conditions; in the first, the users had visual access to the immediate locality of the robot at all times (henceforth, referred to as "Monitor condition") whereas in the second condition, users had no view of the robot or its surroundings (henceforth, "No Monitor condition"). The study is obviously oriented towards the existing prototype but it also attempts to provide general implications for NLIs in goal-directed HRI.

B. *Setup*

A custom Java-based system was developed to support the simulation and the real-time interaction between the participants. It consisted of two applications (interfaces) connected using the TCP/IP protocol over a LAN. The system kept a log of the interaction and, for every message sent or received, the coordinates of the robot at that moment were recorded. During data analysis, we were able to retrace the path taken by the robot with sufficient temporal and spatial accuracy. The interfaces consisted of a map of the town and a dialogue box. The interfaces of the user and wizard are shown in Figures 1 and 2, respectively.

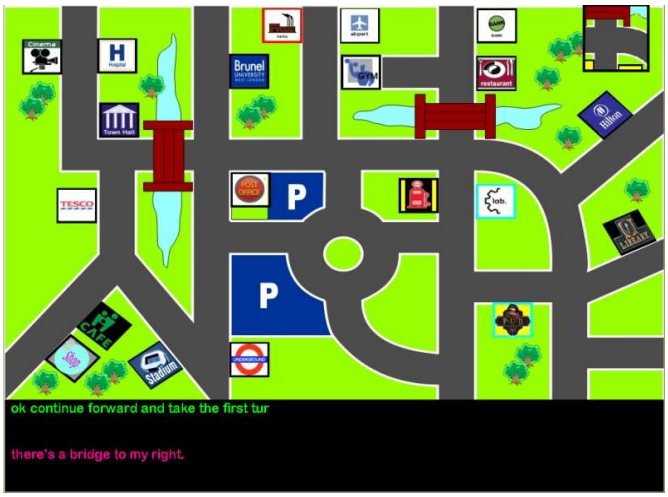

Fig. 1. The user's interface in the Monitor condition. The monitor window on the upper right corner was removed in the No Monitor condition.

The users had a view of the full map. The goal of the current task was shown in red whereas the completed tasks were shown in blue. Similarly to a typical desktop messaging application, they could type messages in the dialogue box and send them to the wizards. The wizards' messages appeared on the lower part of the box (in magenta). Moreover, on the upper right corner of their interfaces, there was a small "monitor", which displayed the robot's surrounding area, but not the robot itself. In other words, the user shared the same visual space with the robot at any point in the interaction. Depending on the experimental condition, the "monitor" feature was added or removed.

The wizard's interface contained a fraction of the map, showing only the surroundings of the robot's current position. The robot was operated by the wizards using the arrow keys on the keyboard. The dialogue box displayed the latest messages by both interlocutors as well as the two previous instructions by the user. Moreover, the existing prototype has the ability to learn routes [15]. Thus, the user could ask the robot to go to a previous location without giving instructions all over again. In order to simulate this functionality in a practical way, after successful completion of a task, a new button appeared on the right side of the wizard's screen which represented the newly learnt route. Therefore, when the user requested to take a known route, the wizard clicked on the corresponding button and the robot automatically executed.

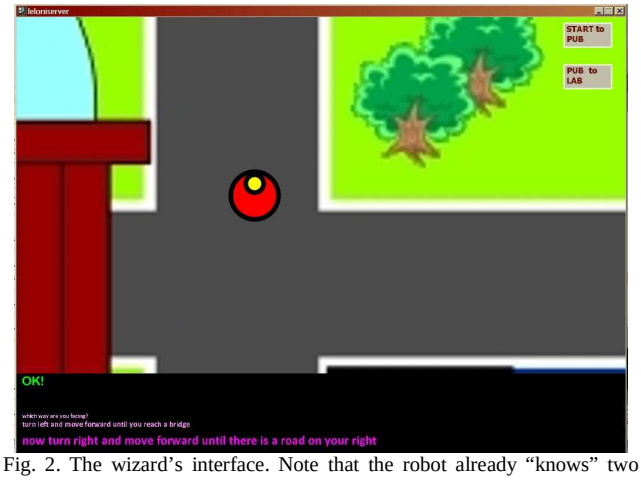

Fig. 2. The wizard's interface. Note that the robot already "knows" two routes.

C. Procedure

A total of 32 students from various departments of the university were recruited (16 users and 16 wizards). Twenty of them were assigned to the Monitor condition and twelve to the No Monitor condition. The allocation of participants to the experimental conditions and roles was random and no computer expertise and other skill were required.

The users and wizards were seated in different rooms equipped with desktop PCs. They received verbal and written instructions. The wizards were given a brief demonstration and time to familiarize with the operation of the interface. Wizards were fully informed about the setup and whether the user was able to see the robot's actions.

The users were made to believe that they would interact directly with a robot, which for practical reasons was the simulated version of the actual robot. They were told that the robot had limited vision, but had advanced mobility and capacity to understand and produce spatial language and it could also remember previous routes. They were asked to open each interaction with "Hello" (which initialised the application of the wizard) and end it with "Goodbye" (which closed both applications). The users were not provided with any examples of what to say. However, it was explicitly requested not to employ "absolute" reference systems, such as "North", "South", "up", "down", which are anyway rarely used in route directions [9], [20]. Moreover, they were asked to take the robot's perspective. Still, in real or simulated settings of HRI, users overwhelmingly do so without being told [20].

Each pair attempted six tasks presented in the same order. In particular, the user had to navigate the robot from the starting point (bottom right on the map) to six designated locations (pub, lab, factory, tube, tesco, shop). The users were free to plan and modify the route as they wished. The destinations were selected to require incrementally more instructions or the use of previously taught routes. Dialogues ran until the user ended them or up to 10-11 minutes (decided on the basis of pilot studies).

It could be argued that any observations from text-based interaction cannot be generalised to spoken dialogue. However, a study in which users navigated a robot using either typed or spoken instructions demonstrated that they employed similar utterances in both modalities [21]. A limitation of the present study could also be the fact that the interface displayed a plan view of the environment whereas in a real-world situation the instructor and follower face three-dimensional objects. Nevertheless, research has shown that 3D concepts play a minor role in linguistic representations and reported no differences in spatial descriptions produced by users in 2D (pictures) and 3D scenarios [22].

D. Dialogue Annotation

The primary annotation of the dialogues was based on the HCRC dialogue act tagging scheme which was designed for navigation dialogues of the HCRC Map Task Corpus [23]. The dialogue acts by the user found in our corpus are shown in Table I below.

Dialogue Act	Description
Instruct	Commands the robot to perform an action.
Explain	States information which has not been elicited by the robot.
Query	Asks the robot any question.
Reply	Any reply to any query.
Clarify	Repeats information which has already been given.
Greet	Hello/Goodbye.

Table I. The tag set used to annotate the user turns.

The dialogue acts tagged as "Instruct" were then classified based on the action-oriented categorisation by [7]. The categories are the following:

1. **Action.** E.g., "Turn left".
2. **Action + Landmark** (landmark, known location, destination). E.g., "Go to the pub.", "Cross the bridge."
3. **Action + Path entity** (road, junction, crossroad). E.g., "Take the road on the right."
4. **Landmark, No action.** E.g., "The lab is on the left".

Finally, following [10], [19], a finer-grained constituent analysis was performed. In particular, we tagged the instructions that contained (a) a projective spatial term, such as "on the right", "on the left", "in front of" etc. , (b) an ordering expression such as "first", "second", "last" etc. and (c) a path-describing term, such as "at", "after", "along", "the end of", "past" etc. Table II below shows an example of an annotated turn. It contains six instructions and the category tag of each is within the square brackets.

Go to the end of the road **[3c]** and turn left **[1]**, go past the bridge **[2c]**, continue straight **[1]** and take the first road on the right **[3ba]**, destination is on the left **[4a]**.

Table II. Example of an annotated turn.

III. Results

The experiments yielded a corpus of 96 dialogues. The dialogues contained 1100 turns by the user (669 and 431 for the Monitor and No Monitor conditions, respectively). This

section reports results of the analysis of the dialogue acts by the user. Then, a component analysis of the instructions and some additional observations are presented. The data were examined both as a whole and by condition.

A. Dialogue Act Analysis

The dialogue acts "Instruct" and "Query" were the most prevalent in the corpus covering 59.32% and 17.96% of all user turns, respectively[1]. However, the analysis revealed striking differences between conditions in terms of occurrence of instruction turns ($t=3.680$, $df=8.521$, $p<0.005$) and questions ($t=-4.270$, $df=5.794$, $p<0.005$). The great majority of user utterances in the Monitor condition were instruction turns (70.58%) whereas questions addressed to the wizard were rare (4.25%). On the other hand, users in the No Monitor condition gave much fewer instructions but issued considerably more queries, with the number of the latter reaching 31.68% (Figure 3).

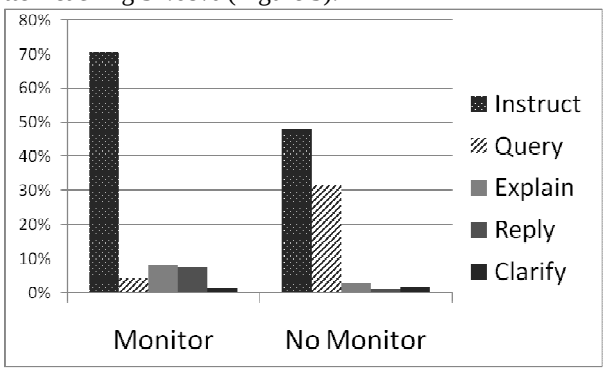

Fig. 3. Occurrence of user dialogue acts in the Monitor and No Monitor conditions.

Qualitative turn-based analysis of the dialogues indicated that when visual information was available, the users could monitor the robot's progress and provide further instructions or corrections at the moment needed (see dialogue in Table III). The interaction was a cycle of the user giving instructions and the wizard executing, reserving additional verbal communication only when repair initiation was necessary. In that case, the users would employ the other dialogue acts, "Explain", "Clarify" and "Reply", to address misunderstandings and clarification requests by the wizard. Surprisingly enough, as reported in [17], there were higher rates of miscommunication in the Monitor condition, which could also account for the larger numbers of "Explain" and "Reply" (Figure 3).

ID	Message	Tag
U1	go to the tube [2]	Instruct
U2	take the left road [3]	Instruct
U3	turn left [1]	Instruct
U4	go straight ahead [1]	Instruct
U5	stop [1]	Instruct
U6	you are at your destination	Explain

Table III. Excerpt of a dialogue in the Monitor condition. ID denotes the speaker (User or Robot). The numbers in the brackets correspond to the instruction categories (see Section II.D).

[1] The "Greet" turns were not considered in the dialogue analysis.

On the other hand, when the robot's action area was not visible, the responsibility for maintaining understanding and assessing the task was balanced between the participants. As exemplified by the dialogue in Table IV, the users had to continuously request and rely on the wizards' verbal descriptions of the environment to determine the status of the task and could not intervene autonomously. The user would query about what the wizard could see, trying to establish a joint focus of attention, perspective and reference; only then attempted to offer further instructions (as in lines 1 and 7 in Table VI). This, of course, led to longer turns ($t=-2.308$, $df=10.388$, $p<0.05$) and task completion times ($t=-2.36$, $df=8.25$, $p<0.05$) for the No Monitor condition. However, task success rates were similar across conditions [17].

ID	Message	Tag
U1	move forward [1], turn right [1], move forward [1], turn left [1] and then stop [1]	Instruct
U2	where are you?	Query
R3	I can't move forward. I am facing a grass field. I can move left or right only.	Reply
U4	What buildings are close to you?	Query
R5	I can't see any buildings.	Reply
U6	turn left [1]	Instruct
U7	can you see any buildings now?	Query
R8	I came to a T junction. I can see a tree.	Reply
U9	turn right [1] and move forward until you get to the next junction [3]	Instruct

Table IV. Excerpt of a dialogue in the No Monitor condition.

B. Instruction Analysis

The corpus contained 561 "Instruct" turns which were decomposed into 798 instruction units.

1) Instruction Units per Turn

As can be seen from the dialogue excerpts, one "Instruct" turn can comprise one or more instructions. There was much inter-subject variability in terms of how many instructions the users embedded in one utterance. Some users preferred issuing one instruction per utterance whereas others provided longer chunks of 3. The average number of instructions per turn in the Monitor and No Monitor condition was 1.54 ($sd=0.55$) and 1.69 ($sd=0.95$), respectively, and no significant effect was observed.

2) Instruction Types

Component analysis of the turns revealed that almost 59.8% of the instructions were simple action descriptions (category 1). On the other hand, users employed landmark references 19.5% of the times (category 2). Path references accounted for the 11.4% of all instructions (category 3). Landmark references without action mostly constituted the final instructions orienting the robots towards the destination and covered 8.1% of the corpus (category 4). The simulated town included a roundabout which was mainly used as a landmark and appeared only in 9 instructions. Figure 4 illustrates the distribution of each type of instruction. The first column denoted "Corpus" corresponds to the data from both experimental conditions as a whole.

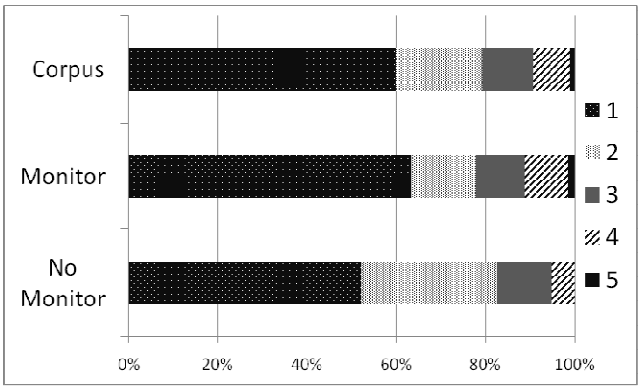

Fig. 4. Use of each instruction category in the whole corpus and in Monitor and No Monitor conditions. 1: Action Descriptions, 2: Action + Landmark References, 3: Action + Path References, 4: Landmark References without Action, 5: References to the Roundabout.

Pair-wise comparison of the conditions revealed an interesting phenomenon. Users in the Monitor condition selected category 1 instructions more frequently than users in the No Monitor condition ($t=2.139$, $df= 11.770$, $p=0.05$). This suggests that when users had no supervision of the robot's actions avoided using simple descriptions of movement. Instead, they employed instructions which were anchored to landmarks such as buildings and bridges ($t=-2.002$, $df=7.452$, $p<0.05$). The use of simpler "two-dimensional" landmarks (roads, junctions) was similar for both groups (Figure 4).

U1	Go straight after tesco [2]
U2	turn right [1]. It's the second building to your left [4].

Table IV. Excerpt of dialogue in the Monitor condition

It was also observed that users in the Monitor condition were more likely to omit boundary information. As in the example in Table IV, the user does not specify up to which point the robot should move forward or whether after taking a turn, it should move forward again until it reaches the destination. On the other hand, when users could not see the robot's execution of the instruction, the level of granularity of their instructions increased. Table V contains the utterances by a user in the No Monitor condition. In the dialogue, the user specified termination points.

U1	can you see the junction on your left?	Query
U2	go to the junction [3]	Instruct
U3	where are you?	Query
U4	turn left [1] and move forward until you reach a bridge [2]	Instruct
U5	move forward until you get to the junction on your left [3]	Instruct
U6	can you see the road on your left?	Query
U7	turn left [1] and move forward until you get to tesco [2]	Instruct

Table V. An excerpt of a dialogue in the No Monitor condition. The wizards' responses are removed.

3) Use of Deixis

Deictic expressions such as "this", "that", "here" and "there" are used for indexing entities in the local surroundings. They are generally preferred by speakers, as they substitute for longer referring expressions that are based on spatial relations like "left", "right", "front" etc. [14]. However, they require both conversational partners to establish that these entities are in their joint attention. The aforementioned studies in human communication (Section I.C) showed that a shared visual space increases the use of these expressions. Thus, in the context of this study, users in the Monitor condition were expected to make extensive use of instructions such as "take this road" or "turn left here".

Analysis of the users' instructions in our corpus did not provide support for this hypothesis. In fact, the use of these elements was very rare (7 instances in 798 instructions). In the Monitor condition, 8 out of 10 users never used them and they appeared more than twice in the instructions of just one user. In the No Monitor condition, there was only one instance. Therefore, due to the small numbers and individual variability, it is not possible to infer that visual information had an effect on the use of deixis. Nevertheless, it could be concluded that users do not generally opt for underspecified deictic expressions to navigate the robot.

4) Projective, Path-describing and Ordering Terms

The number of projective, path-describing and ordering terms was measured. Projective and path-describing expressions were used in half of the instructions in the corpus (47.2% and 56.3%, respectively). Ordering terms appeared in 3.85% of the instructions. The occurrence of these elements in relation to the instruction categories was also considered in the analysis. The results show that when users referred to landmarks (second category), they did not generally incorporate any of these elements (86%). However, this phenomenon could be quite experiment-specific. The users frequently used previously taught routes, so they would request the robot to re-take a route with a simple instruction, such as "go to the tube." The rest of the landmark references included a path-describing term. Regarding references to path entities (roads, junctions, crossroads), the large majority of these instructions (71%) were further specified by a path-describing terms, whereas projective terms were found in 15% of them. Simple path references such as "go to the junction" also accounted for about 14%. Furthermore, almost all instructions in the fourth category contained projective terms. Hence, users tended to provide the final instruction with utterances such as "the pub is on your left". In our corpus the users did not generally combine projective and path-describing terms in one instruction. Figure 5 summarises these findings and illustrates the distribution of projective and path-describing elements in terms of instruction category.

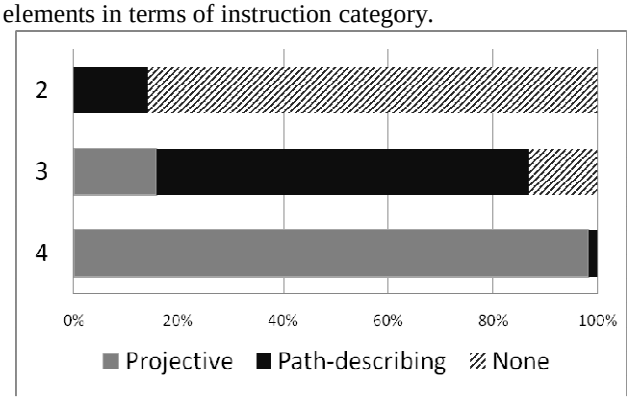

Fig. 5. Use of Projective and Path-describing terms for each instruction type

5) Comparison with the IBL Project

Our current work aims to build upon the IBL project [18]. For that project, a corpus collection was performed in order to define a lexicon for tuning the speech recognition engine and a functional vocabulary with primitive navigation actions. The route instructions in their corpus yielded 15 primitive procedures (e.g., TAKE THE [<number>] turn [(left | right)] | [(before | after | at) <landmark>]) which were then pre-programmed in the robot. The setup of the study involved 24 subjects giving short or long spoken instructions in a monologic setting or in a simplified dialogue with an operator. The route instructions were not executed during the experiment.

The experiments described here follow that setup in only a few respects; that is, the task was navigation in a town and the users could use previously taught routes. The corpus of our study produced route descriptions derived from text-based interaction in a situated human-robot collaborative setting, by also varying the complexity of the task. However, from a preliminary analysis, it can be concluded that the navigation actions in our corpus are consistent with and replicate the set of primitive procedures as defined by the IBL project. The instructions in our corpus are currently mapped to the set of primitives and quantitative analysis is in progress to estimate their respective frequencies. This observation supports the argument that spoken and typed route instructions are structurally and semantically similar [21]. But most importantly, it could suggest that the primitive actions and approach of the IBL project are not domain-specific but can extend and be relevant for various applications. With regard to our current work, it indicates that there is no need to add more machine procedures to the robot manager, but, instead, focus our attention on enhancing the dialogue manager of the system.

C. Dialogue Synchronisation

Robots are usually built on agent-based architectures. However, since all agents (including the human agent) can send information at any time, distributed systems often face serious problems of synchronisation [24]. So, in a typical scenario- also observed in our data- the robot receives a new instruction before having processed the previous one. The new instruction would probably be interpreted within the "old" context leading to a wrong execution. Lack of synchronisation also occurs at the turn-taking level. Namely, the robot and user's utterances may overlap or the user "barges-in" while the robot is speaking. Moreover, untimely feedback and clarification requests by the robot can have a severe impact on the dialogue [19].

The experimental setup of this study enabled us to observe the effect of synchronisation problems on the interaction. The messages were formulated in a private window and were transmitted when the participants pressed "enter". Thus, they only became aware that the other participant had been addressing them when they received the message in full. As a consequence, 5.2% of all turns overlapped or were delayed. The users typically ignored such messages and proceeded with the dialogue. Occasionally, however, an ill-timed request by the wizard would cause the user to give erroneous instructions. In the example in Table VI below, the messages in lines 2 and 3 are sent simultaneously. In line 4, the wizard has already executed, but the user still perceived the wizard's request as relevant and repeated the instruction. If instructions are "stacked" for execution, this would lead the robot to execute the same instruction twice (as often did the wizards).

ID	(x,y@time)	Message
U1	(784,441@10:27:25)	turn left again
U2	(727,560@10:27:52)	ok, turn right now
R3	(727,560@10:27:52)	Where now?
U4	(585,424@10:28:43)	turn right

Table VI. Excerpt of a dialogue in the Monitor condition. (x,y@time) show the coordinates of the robot (on a 1024x600 pixel map) and the time the corresponding message was sent.

IV. Discussion and Further Work

The findings presented in Sections III.A and III.B.2 suggest that the pairs coordinated differently and users employed different strategies depending on whether monitoring was possible. This effect was observed in the dialogue acts of the user and it extended deeper into the formation of instructions. Namely, when users could not see what the robot saw and did, they usually provided instructions clearly stating boundaries and actions. Furthermore, after having established joint reference, they integrated these landmarks into their descriptions. Landmark references serve as cues for (re-)orientation and are used to solve or prevent problems [9], and were prevalent in the No Monitor condition. References to path landmarks were found as frequently in both groups. However, compared to landmarks such as bridges and buildings, path descriptions hold lower information value in navigation. By contrast, several users in the Monitor condition relied entirely on underspecified purely spatial instructions which often lacked boundary information (Section III.B.2). This observation relates to the results discussed in Section III.A and suggests that since in the Monitor condition, the dialogue and execution were synchronous, the user was able to provide the next instruction with temporal precision, at the moment in which the wizard was observed to have completed the previous instruction. Thus, the wizards assumed that "move forward" means "move forward, until I tell you to stop or give you a new instruction." In fact, the command "stop" regularly appeared as an instruction (without intention to correct an error) in the turns of three users in the Monitor condition (reaching 55% for one of them). Robots lack such inferential capacity. It was additionally observed that although users in both conditions could not know the robot's orientation at all times (the monitor did not display the robot), users in the No Monitor condition were more inclined to find out before giving directions (as in line 3 in Table VII below).

ID	Message	Tag
U1	where are you?	Query
R2	in front of the lab	Reply
U3	are you facing the lab?	Query
R4	Yes	Reply
U5	can you see the junction on your left?	Query
R6	I can only see a part of a junction and there is a building behind me	Reply
U7	Go to the junction [3]	Instruct

Table VII. Excerpt of dialogue in the No Monitor condition.

Our findings are consistent with the collaborative view of human communication [5]. The users formulated their instructions based on assumptions about the information the robot needed. Thus, by seeing the robot performing the task, the users could easily confirm their hypotheses and use linguistic shortcuts and simpler constructs. But the principle of least effort is always balanced with the need to ensure understanding, so in the No Monitor condition, in which the users were not sure if they see the same spatial positions as the robot, they had to provide explicit and more elaborate instructions. This finding suggests that spatial language should not be studied in isolation, but in realistic, dialogic settings.

It is also necessary to draw a distinction between visual and full physical co-presence. In the former, the partners maintain a common visual space whereas in the latter, spatial relations between the interlocutors, task-relevant objects and the wider environment are attended [11]. In our experimental setup, the users did not employ simple, underspecified deixis. It could be assumed, however, that in fully situated interactions, the use of these expressions will be pervasive. Therefore, dialogue strategies to address such linguistic elements should be integrated in the dialogue manager of robots that are destined to interact with users within the same space. On the other hand, in remotely-controlled or (semi-) autonomous robots such ability could be less essential.

The results provide additional implications for HRI. They indicate that the route descriptions employed by users who can monitor the robot can be less detailed and precise. Consequently, the demands for spatial reasoning increase for a collocated robot. However, a robot which does not share its visual space with the user faces another challenge; when monitoring was restrained, the users continuously requested information about the current location of the robot. A "human" robot was certainly able to provide rich descriptions of its surroundings. Providing effective feedback is crucial for task-oriented interactions, and especially in the dynamic setting of HRI, in which the user's instructions can be incomplete or outdated. However, this is not a trivial task; the architecture of a speech-enabled mobile robot involves several components typically divided in two modules, one for interpreting and generating language and one for processing and executing the actions. Situated dialogue entails instantaneous synchronisation and updating of these modules to include a continuous influx of information. Thus, clarification requests and feedback need to be provided with high temporal accuracy, or else, they could impair the interaction and lead to confusion and errors (see Section III.C). Furthermore, providing redundant feedback compromises the "naturalness" and efficiency of the interaction. Our empirical results also argue that the execution of the task is often sufficient feedback by itself [4]. Therefore, when and what kind of feedback to provide should be determined by a criterion that draws on several knowledge sources and is updated both within and between sessions [25]. These sources could be the dialogue history (e.g., how many times in the dialogue so far the robot and user have initiated repair), model of the environment (e.g., is the robot at home, outdoors or at a crowded workplace) and the task (e.g., is the route well-known, what are the consequences of errors). As part of our future work, we will focus on issues pertaining to the implementation of such functionality.

According to a study on spatial descriptions [26], clarification requests have a direct effect on the processes of coordination. They observed that as the dialogue progresses, the pairs converge in the use of more complex and efficient spatial descriptions. However, after clarification sub-dialogues, the instructor shifts to more conservative descriptions. These insights coming from human communication present interesting questions and rich opportunities for investigation in HRI. In particular, there has been considerably less research on how the linguistic content of the users' instructions change over the course of the dialogue and how repair initiations by the robot would affect the choices of the user. It has been observed that users tend to recycle utterances which were previously successful [20]. However, the robot in that study had a limited and fixed response repertoire. In our research, robot navigation is strictly viewed as a bilateral process. The current study demonstrated how users adapted their linguistic behaviour according to the demands of the experimental condition. Thus, our next step is to examine the route instructions within the course of the dialogue; that is to say, whether the users revised and adapted their strategies in response to particular robot utterances and in the presence of miscommunication. Last, it would be interesting to determine whether certain user strategies (as primed by previous robot responses) are more efficient in terms of recovery from error.

One of the most challenging endeavours in the design process of a NLI for a robot is to "enact" a HRI scenario that permits natural language and behaviour by the users but is also realistic and supports the future implementation of the system. The present study recreates an urban navigation scenario in which non-experienced users interact and teach a mobile robot. It involves two configurations of supervised and unsupervised interaction and is primarily explorative. The current trend in linguistic and robotics research is the joint investigation of spatial language and dialogue [27]. Thus, in our study, route instructions are collected in a dialogic situation. In this setting, information and

understanding are continuously monitored and revised. The naturalness of the data was assured by the lack of an informed "confederate" and a dialogue script. The results were interpreted as a corpus of route instructions following analytic paradigms established by previous research. These results seem to be aligned with and extend findings from a range of disciplines (human communication and spatial cognition) and from various application areas and tasks in HRI.

V. CONCLUSION

This paper describes the collection and analysis of route instructions in a simulated HRI study. It illuminates patterns of linguistic behaviour of the users, also resonating with findings from studies in human collaborative behaviour and HRI. On one hand, the analysis of the data provides support for the "action-oriented" computational models of spatial language [18], [28], which treat instructions as physical actions moving the agent within space. On the other hand, this study collected route instructions as they emerged from dynamic interaction with the recipient. As expected, route instructions were often problematic but the participants managed to coordinate in the information and semantic levels and completed the tasks. It also became evident that even minor synchronisation problems can have a serious effect on the interaction. The next step in our research is to create a dialogue model based on these findings. The focus of the work will be on enriching the dialogue manager of the robot with "human-inspired" mechanisms to negotiate insufficient information and help the user provide instructions that it can interpret. Such framework could produce insights that extend beyond dialogue-based navigation and are applicable to different domains of goal-oriented HRI.

REFERENCES

[1] S. Thrun, "Toward a framework for human-robot interaction", *Human-Computer Interaction*, vol. 19, no.1, pp. 1-8, 2004.

[2] M. Gabsdil. "Clarification in spoken dialogue systems", in Proc. of 2003 AAAI Spring Symposium Workshop on Natural Language Generation in Spoken and Written Dialogue, Stanford, USA. 2003.

[3] G. Bugmann, "Spoken interfaces to service robots: Open problems," in *Proc. AISB2005*, Hatfield, UK, 2005, pp. 18–22.

[4] A. Green, B. Wrede, K. Severinson-Eklundh, and S. Li, "Integrating miscommunication analysis in natural language interface design for a service robot," in *2006 Proc. IEEE/RSJ, Int. Conf. on Intelligent Robots and Systems*. Beijing, China.

[5] H.H. Clark and C. R. Marshall, "Definite reference and mutual knowledge", in *Elements of Discourse Understanding*, B. L.Webber, A. K. Joshi, and I. A. Sag, Eds. Cambridge, UK: Cambridge University Press, 1981, pp.10-63.

[6] T. Tenbrink, "Communicative aspects of Human-Robot Interaction," in *Languages in Development*, H. Metslang and M. Rannut, Eds. Lincom Europa. 2003.

[7] M. Denis, "The description of routes: A cognitive approach to the production of spatial discourse," *Current Psychology of Cognition*, vol.16, no.4, pp.409-458. 1997.

[8] M. MacMahon, B. Stankiewicz, and B. Kuipers, "Walk the talk: Connecting language, knowledge, and action in route instructions", in *Proc. AAAI-2006*, Boston, MA

[9] M. Denis, F. Pazzaglia, C. Cornoldi, and L. Bertolo, "Spatial discourse and navigation: An analysis of route directions in the city of venice," *Applied Cognitive Psychology*, vol.13, no.2, pp.145–174, 1999.

[10] T. Tenbrink, V. Maiseyenka, and R. Moratz, "Spatial reference in simulated human-robot interaction involving intrinsically oriented objects", in *Proc. Symposium Spatial Reasoning and Communication at AISB'07*, Newcastle upon Tyne, UK.

[11] R. E. Kraut, S. R. Fussell, and J. Siegel, "Visual information as a conversational resource in collaborative physical tasks," *Human Computer Interaction*, 18, 13-49, 2003.

[12] H.H. Clark and M. A. Krych, "Speaking while monitoring addressees for understanding". *Memory & Language J.*, vol. 50, pp. 62-81, 2004.

[13] S. E. Brennan, "How conversation is shaped by visual and spoken evidence," in *Approaches to Studying World-situated Language Use: Bridging the Language-as-product and Language-action Traditions*, J. Trueswell and M. Tanenhaus. Eds. Cambridge, MA: MIT Press, 2005, pp. 95-129.

[14] D. Gergle, R. E. Kraut, and S.E. Fussell, "Language efficiency and visual technology: Minimizing collaborative effort with visual information," *Journal of Language and Social Psychology*, vol. 23, no. 4, pp. 491-517, Thousand Oaks, CA: Sage Publications. 2004.

[15] S. Lauria, G. Bugmann, T. Kyriacou, J. Bos and E. Klein, "Training personal robots using natural language instruction." *IEEE Intelligent Systems*, pp.38–45. 2001.

[16] T. Koulouri and S. Lauria, "A WOz framework for exploring miscommunication in HRI", in *Proc. AISB Symposium on New Frontiers in Human-Robot Interaction*. Edinburgh, UK. 2009.

[17] T. Koulouri and S. Lauria, "Exploring miscommunication and collaborative behaviour in Human-Robot Interaction", in *Proc. SIGdial09*. London, UK. 2009.

[18] G. Bugmann, S. Lauria, T. Kyriacou, E. Klein, J. Bos, J., and K. Coventry, "Using verbal instruction for route learning: Instruction analysis," in *Proc. TIMR01-Towards Intelligent Mobile Robots*, Manchester. 2001.

[19] T. Tenbrink and S.Hui, "Negotiating spatial goals with a wheelchair," in *Proc. 8th SIGdial Workshop*, Antwerp. 2007.

[20] R. Moratz and T. Tenbrink, "Spatial reference in linguistic human-robot interaction: Iterative, empirically supported development of a model of projective relations," *Spatial Cognition and Computation*, vol.6, no.1, pp. 63–106, 2006.

[21] R. Moratz and T. Tenbrink, "Instruction modes for joint spatial reference between naive users and a mobile robot," in *Proc. RISSP IEEE*, Changsha, China. Oct. 2003.

[22] T. Tenbrink, "Methods for analyzing natural discourse: Investigating spatial language in HRI vs. in a no-feedback web study," in *Proc. Dagstuhl Seminar on Spatial Cognition: Specialisation and Integration*, Germany. 2007.

[23] J. Carletta, A. Isard, S. Isard, J. Kowtko, G. Doherty-Sneddon and A. H. Anderson, "HCRC dialogue structure coding manual", (HCRC/TR-82), University of Edinburgh. 1996.

[24] N. Baylock, J. Allen, and G. Ferguson, "Synchronization in an asynchronous agent-based architecture for dialogue systems," in: *Proc. 3rd SIGDial*, Philadelphia, USA, pp. 1–10, 2002.

[25] S. E. Brennan and E. A. Hulteen, "Interaction and feedback in a spoken language system: A theoretical framework," *Knowledge-Based Systems*, vol. 8, pp. 143-151. 1995.

[26] G. Mills, and P. G. T. Healey. "Clarifying spatial descriptions: Local and global effects on semantic co-ordination," in *Proc. 10th Workshop on the Semantics and Pragmatics of Dialogue*. 2006.

[27] K. Coventry, T. Tenbrink, and J. Bateman, "Spatial language and dialogue: Navigating the domain," in *Spatial Language and Dialogue*, K. Coventry, T. Tenbrink, and J. Bateman. Eds. Oxford University Press. 2009, pp. 1-8.

[28] H. Shi, C. Mandel, R.J. Ross, "Interpreting route instructions as qualitative spatial actions," in *Spatial Cognition V*, T. Barkowsky et al. Eds, Berlin Heidelberg: Springer, 2007, pp. 327-345.