Automatic Verbal Depiction of a Brick Assembly for a Robot Instructing Humans

Rami Younes, Gérard Bailly, Damien Pellier, Frédéric Elisei

Automatic Verbal Depiction of a Brick Assembly for a Robot Instructing Humans

[1,2]Rami Younes, [1]Gérard Bailly, [2]Damien Pellier, [1]Frédéric Elisei
[1] Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France
{firstname.lastname}@gipsa-lab.grenoble-inp.fr
[2]Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France
{firstname.lastname}@univ-grenoble-alpes.fr

Abstract

Verbal and nonverbal communication skills are essential for human-robot interaction, in particular when the agents are involved in a shared task. We address the specific situation where the robot is the only agent knowing about both the plan and the goal of the task, and has to instruct the human partners. The case study is a brick assembly. We here describe a multi-layered verbal depictor whose semantic, syntactic, and lexical settings have been collected and evaluated via crowdsourcing. One crowdsourced experiment involves a robot-instructed pick-and-place task. We show that implicitly referring to achieved subgoals (stairs, pillars, etc) increases the performance of human partners.

1 Introduction

Task-oriented interactions between systems and humans, in order to achieve a common goal, are present in many applications. For instance, receiving directions via GPS is a task-oriented communication with the instructions being delivered visually and verbally (Belvin et al., 2001). Similarly, robots have been used to give directions (Bohus et al., 2014), describe its route experience (Rosenthal et al., 2016; Perera et al., 2016) or instruct students in a tutorial class Gomez et al. (2015). In the later work (see Figure 1), a robot helps two participants to perform a jigsaw assembly task using verbal and non-verbal communication.

Figure 1: Face-to-face interaction on a Jigsaw reassembly task with an Icub robot (right) acting as the instructor for two students (left). From (Gomez et al., 2015).

The rise of social robots endowed with verbal, co-verbal and non-verbal communication capabilities, now raises the question from the robotic point of view. How a robot and a human can communicate to achieve a common goal and share plans, involving manipulating objects in their common working space?

In this paper, we study this problem by focusing on verbal communication. Indeed, a verbal description of how the task is to be done is a more effective way of communicating objectives than non-verbal descriptions: not only does it improve task performance but also gives rise to more compliance and better mutual adaptation (see Nikolaidis et al., 2017). More precisely, we propose to explore the impact of the verbalization strategy in an extreme case where the robot is the only agent that knows the plan and the goal, and human coworkers are awaiting instructions planned by the robot to achieve the goal. Note that we limit here the number of human partners to one: the opportunistic allocation of tasks between available coworkers will be addressed in a following paper.

When using verbalisation as the main means of communication, an important question is to see how the style, i.e. saying the same thing in different ways, quoted as the *verbalization space* by (Karlgren, 2000), in which the instructions are being delivered, can affect the execution of the task, especially when communicating complex tasks.

In this context, our contribution is threefold:

Styles: we test four different styles on an assembly task (see Figure 2) that offers a large verbalization space, i.e. many stylistic dimensions including choice of geometric relations between bricks, of syntactic and lexical descriptions, etc. One primary style parameter is the use of *context*. By context, we mean implicit referencing to elements of the environment that go beyond the previous and current actions. We compare two AI-generated styles

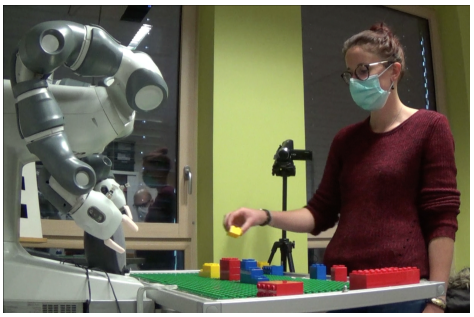

Figure 2: Face-to-face interaction on a LEGO™ assembly task with YUMI acting as the instructor.

(inclusion vs. dismissal of the use of context) with the lowest vs. highest human instructions in terms of: time-to-complete, comprehension/complexity of the instruction, and efficiency/effectiveness of the task completion.

Architecture: we propose a robotic control architecture and its sensorimotor capabilities for task-oriented human interaction. We focus on two key components: (a) the planner and (b) the verbalizer. While the former takes decisions on what to do next, the verbalizer puts each elementary instruction into words for a text-to-speech synthesizer.

Evaluation: we propose an evaluation framework based on a series of three crowdsourced experiments for: (a) collecting and (b) scoring human verbalizations for parameterizing our flexible verbalizer as well as (c) assessing and evaluating the performance of human vs. automatic verbalizations on actual task assembly.

This paper is organized as follows: section 2 contains related work for task-oriented communication and plan description; section 3 describes the overall architecture of our control model, with a closer look on the planner and the verbalizer with its different layers (fig. 6); section (4) introduces the three experiments used to parametrize the verbalizer; section 4.1 presents the results of the first two web-based experiments used for data collection and assessment of human verbalizations; finally, section 4.2 presents the setup and results of the last web-based experiment used to validate the efficiency of the set of rules in our automatic verbalizer and to compare it with human verbalization.

2 Related work

Verbal communication of plans has been used in a large variety of Human-Robot Interaction (HRI) scenarios. They mainly vary along four main categories: (1) task type (e.g. commentating, instructing, navigation). (2) perception capabilities (auditive/visual sensors). (3) style and its use in sentence generation (e.g. information tagging). (4) role (e.g. receptionist, instructor, navigator).

Most HRI tasks require some form of communication. It could be used to describe what happens in the scene: Veloso et al. (2008) presented Rocco, a fully automated RoboCup (Kitano et al., 1997) commentator, aiming at generating real-time summaries of the actions in the games (Voelz et al., 1998). For navigation tasks, Rosenthal et al. (2016); Perera et al. (2016) presented algorithms for generating routing narratives with varying parametrized styles. Belvin et al. (2001) presented a real-time spoken language navigation system able to respond to natural conversational queries. The queries were mainly regarding details of a step in the route. However, responses were generated using simple pre-written "holly sentences" filled with variables extracted from the plan. For assembly tasks, the 'SHRDLU' system (Winograd, 1972, 1974) is quite inspiring: the task focused on manipulating blocks with a robot arm on the basis of the user's textual input. The system translates the user's input into procedures to move the blocks and question the scene. Our work exchanges the roles of the agents: our robot instructs human agents verbally. Fiore et al. (2014) also shares some similarities with our work, i.e. verbalising the actions in the plan for the user as well as explaining which actions should be executed and in what order. However, the task they chose does not require the same level of precision – and their focus was not on verbalization. Finally, the Robert system (Behnke et al., 2020), installed on Bosch equipment, provides its user with a step-by-step instruction (on a screen using text, images, voice and videos) detailing how to complete a given DIY project successfully. Similarly to us, the sequence of instructions (plan) is obtained using HTN planning. They also added a new feature to perceive the scene using connected tools (sensors), enabling the system to check whether the user is performing the project's steps correctly and to provide help in the case of failure.

Zhu et al. (2017) proposed a verbalization system able to generate explanations for navigation as well as grasping and manipulation tasks (pick-and-place kitchen scenario). They used pre-written sentence templates. Canal et al. (2021) proposed PlanVerb, a domain and planner-independent method

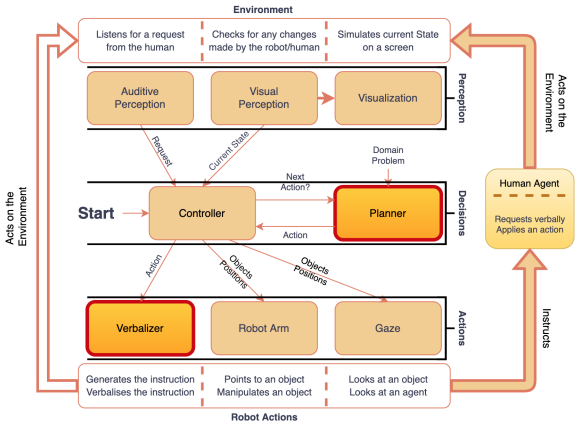

Figure 3: Outer part shows the Human-robot Interaction. The inner part shows the multimodal Architecture. The two components involved in the paper are highlighted .

for the verbalization of task plans based on semantic information tagging of the actions and predicates in the domain (for both PDDL and RDDL). Several works showed the importance of having different styles (Aires et al., 2004; Miehle et al., 2018a; MacFadden et al., 2003). In line with Voelz et al. (1998); Veloso et al. (2008), our verbal generation framework relies on crowdsourcing experiments, for both defining the different styles of the instructions and assessing their efficiency.

Verbal communication has been used in a large variety of situations. Gockley et al. (2005) proposed a robot receptionist with pre-written storylines. Their focus was on long-term interactions with a robot that exhibits personality and character. For their robot bartender, Petrick and Foster (2013) construct plans with tasks, dialogue, and social actions. They advocate for a stronger link between planning and language.

3 The architecture

Figure 3 shows the overall architecture of our HRI system monitoring the interaction between the agents (humans and robot) and the working environment. While no single architecture has proven to be best for all applications, layered architectures have proven to be increasingly popular, due to their flexibility and ability to operate at multiple levels of abstraction simultaneously (Kortenkamp et al., 2016). Similarly to what can be found in (Alami et al., 1998), our robot's architecture can be divided into three levels: perception, decision making and action. With different robots, capabilities change and so do their perception/action modalities. This architecture allows us to add/remove

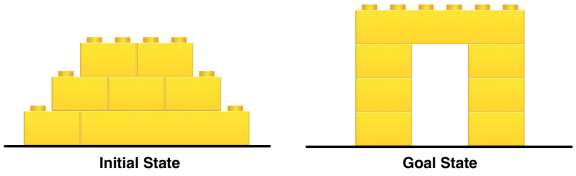

Figure 4: An example where the task is to build a LEGO™ arch – Hierarchical decomposition in fig. 5

modalities, in order to cope with both industrial (e.g. no gaze/head) and humanoid robots (e.g. no grippers).

The perception level is in charge of capturing the current state of the environment as well as the agents acting on it, e.g. analyzing verbal requests coming from the human agent and all changes happening in the working environment. The action level takes charge of all actions towards the environment (e.g robot moving around to pick an object) and agents (e.g. coordinated gaze, speech and pointing to attract partners' attention). The controller is responsible for orchestrating action/perception loops according to the current objective given on request by the planner. In particular, the controller is in charge of monitoring the addressee's activity when processing the robot's instruction, such as on-line attention, task comprehension and correct execution. This includes the chunking or repetition of the instruction if necessary.

It all starts with the controller that receives a "go" signal and requests the first action from the planner. Provided the requested information, the controller (via the action modules) either applies the action or instructs the human agent to apply it. The environment is modified, the controller perceives the updated current state, and the loop continues until the planner deems this task as completed.

The following subsections detail the key modules for the generation of verbal instructions: the planner and the verbalizer.

3.1 Planner

The planner takes as input a domain that contains a logical description of the actions, an initial state obtained from the perception layer and an objective and outputs a sequence of actions (the plan/solution) in order to reach the objective. The initial state and the objective are described as a set of logical propositions.

The first advantage of using a planner is (a) being able to scale to other types of tasks (e.g. assembly,

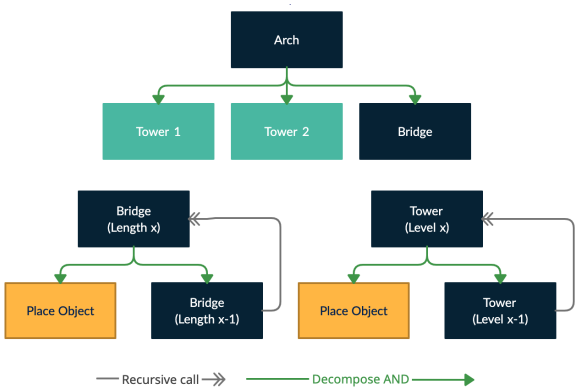

Figure 5: Hierarchical decomposition of an arch into towers/pillars and a bridge/beam – Example in fig. 4

real-world applications, navigation, etc.), by adapting the domain and the problem to the new task; (b) allowing the robot to autonomously adapt and replan when observed actions differ from expected ones. Our planner has two important properties: it performs Hierarchical Planning (Pellier and Fiorino, 2018), i.e. the goal task can be divided into subtasks, and Partial order planning, i.e. some actions can be executed in parallel as long as they satisfy applicability constraints. Hierarchical Planning allows us to specify subtasks, name them, and use these names in the verbalizer. Partial ordering means that the sequence of actions does not need to be fixed and some actions, while satisfying applicability constraints, can be executed in a different order. Partial ordering eases task description and offers more flexibility while executing the plan or verbalizing it. For instance, building an *arch* can be decomposed into building two *pillars* and a *beam*. Each pillar can be constructed by different workers but the beam assembly requires pillars to be finished.

This provides a plan that is almost identical to how a human would plan to build an arch. Thus, enabling the planning system to provide context about the plan as well as some explanation regarding its decisions in the plan. One might argue that context may not be crucial when giving an instruction. However, when dealing with a complex/important task, the addition of hierarchical decomposition into subtasks can help with assigning separate subtasks to different users, or giving a clear explanation to why we are applying a certain action. Our focus for using hierarchy is to give context to help remove ambiguity from instructions, and reduce the number of errors and needed time to complete the task.

The planner takes into account other constraints such as visibility (cannot perform an action if it prohibits you from seeing a later action), applicability (cannot apply what is inapplicable in a given state), and hierarchical constraints (best to finish all actions of a subtask before starting another one). The planner module also provides vital contextual information for the completion of that action.

3.2 Verbalizer

We communicate the instructions via verbalization. The aforementioned verbalizer has multiple parameterized layers (see Figure 6), each shaping one aspect of the message:

The depictor takes charge of all geometric aspects which are vital for completing an action (e.g. 3D position, orientation). This is where business ontologies are hosted (presently, what characterizes a pillar, steps, windows, walls, etc)

The semantic generator focuses on the context, which is in our case giving a hierarchical explanation of where and why we are applying a certain action (e.g. "To finish the red tower").

The syntactic generator focuses on the syntax (i.e. arranging the words and phrases to create well-formed sentences).

The realizer generates the final sentence from the syntactic tree

The text-to-speech system converts the text into audiovisual signals

Style parameters condition each layer so that to be able to adapt communication to the task difficulty and workers' competence, with the objective to improve performance – e.g fewer mistakes and faster completion time of the instruction (Cassell and Bickmore, 2003; Forbes-Riley et al., 2008; Stenchikova and Stent, 2007; Reitter et al., 2006; Mairesse and Walker, 2010; Miehle et al., 2018b) – as well as cognitive load – e.g better recall of the task and processing capacity.

This multi-layer architecture was chosen in order to separate the different skills of the verbalizer, therefore constraining interventions when extending its capabilities. We first discuss each layer separately and spot differences between car navigation vs. assembly task.

3.2.1 Depictor

The depictor handles here agents, objects, and predicates that populate a particular domain: here 3D arrangement of objects. It contains all necessary

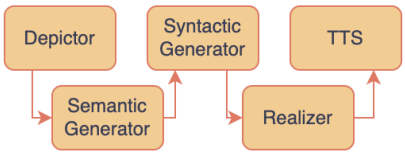

Figure 6: The different layers of the Verbalizer.

properties of these elements such as dimension, relative position, color of objects, sets of objects (e.g. pillars, arches, etc), their relations as well as possible actions (e.g. placing, straddling, sticking, etc). Relative positioning is used by the verbalizer to describe *where to place what, how and why*, e.g. relative to the closest LEGO™ object, or the last placed one. It can be relative to more than one object, or even a structure. For example: it enables the generation of *"To finish the south pillar, put another red brick on top of the previous red brick"*.

It takes as input the action to be performed, the observed scene, as well as the goal, i.e. hierarchical information of the desired final arrangement; and outputs all possible spatial descriptions of the next action using metric, directional and topological operators as seen in (Borrmann and Rank, 2009). For instance, "*Stick* a blue cube, *East of* the previous cube then *move it two slots to the South*". *Stick* translates to the blue cube touching the previous one which is a topological operator. *East of* is a directional operator. Lastly, *move it two slots to the South* is a metric putting forward the distance between the two objects in a certain direction.

3.2.2 Semantic generator

Given all possible actions delivered by the depictor, this layer filters/prioritizes the output list of the depictor according to the style policy: efficiency of the description in terms of positioning (e.g. use of centering, alignments), displacements, use of context, etc. When a chosen description is potentially ambiguous, it may add to the corresponding action extra verification(s).

The context is coming from the hierarchy of the tasks delivered by the planner (e.g. Without context: *"Put a red brick on top of the previous red brick"*. vs. with context: *"To finish the red tower, put a red brick on top of the previous red brick"*). Note that it is also responsible for adding information on the addressee (who should perform the task) and the task (e.g. explaining what it consists of and why this action is triggered, e.g. *"Let's start building an arch starting with its north pile!*).

3.2.3 Syntactic generator

The syntactic generator is responsible for building a syntactic tree with proper verbal constructs, names of objects, etc

Finally, in this layer, we have the option of either including all of the information (verbose) or omitting any redundant information as well as including pronominalisation, all while preserving the unicity of the task. (concise). Verbose: *"Put a red brick on top of the previous red brick"*. Concise: *"Put another one on top"*. The verbose option is straightforward and simply includes everything there is to know about the action. When applying the concise option, in order to remove redundant information, we need to consider what was previously manipulated by the human agent (i.e. LEGO™ type, color, orientation, task).

3.2.4 Realizer

It converts a syntactic tree into sentences. We used the jsRealB (Molins and Lapalme, 2015), that can handle both English and French (fig. 7).

3.2.5 Text To Speech

We currently use the macOS TTS. One issue that we have encountered is problematic mispronunciations in French (in particular handling homographs, liaisons, etc). The current TTS also does not allow to change the intonation in case we decided to manipulate the style of speech (e.g. instructing an order or an astonishment), nor can we include pauses to include coordination with gesture and gaze. Future work will include the use of a different TTS which allows controlling expressivity and rhythm as well as adding emphasis on certain parts of the text.

Going back to the idea of parametrizing the module for another task, a spatial task to be precise, a navigation task for instance. Aside from the necessary updates in the domain and problem files of the planner module corresponding to the new task at hand, some changes need to occur in the first two layers of the verbalizer. The depictor would still generate the semantic depiction and the relative positions between the objects, however, we would need to introduce the newly added different types of objects from the new environment (e.g. immovable obstacles, roads, traffic lights) and actions ("turn", "cross", "look", etc) as well as some information about the role and the link between these objects. The semantic generator would have the same objective as well, however, changes might

Figure 7: English and French realization using jsRealB

be required to accommodate with the type of information that needs to be transmitted. As for the rest of the layers, no particular update is required since it only concerns the styling, generation and utterance of the sentence.

It is important to mention that previous work, such as (Dogan et al., 2020), have shown that including perspective-taking helps reduce time and error. However, their work also mentions that the use of 'in front' and 'behind' did generate some ambiguity and caused more time and errors when applying a task. In our work, the use of left and right could have been easily used instead of East and West since we know the user's position with respect to the environment. However, we decided to use conventional directions (i.e. North East West South) instead of using perspective taking, (1) to ensure the absence of any ambiguity and (2) since this formulation can be used to instruct multiple users having different perspectives.

The following sections include the three web-based experiments that we conducted for finding out which (depiction/verbalization parameters) combination offers the best reduction of errors to complete the assembly.

4 Experiments

The purpose of the first two experiments is to provide us with a ground-truth corpus of verbal descriptions and obtain the highest and lowest ranked human exemplars for giving an instruction. The third experiment introduces an assembly task in order to test the efficiency of AI-generated instructions with reference to the natural ones.

Following the spatial representation — using metric, directional and topological operators as well as using multiple types of object references (point/corner, line/axis . . .) (Borrmann and Rank, 2009) — we chose a set of elementary actions which (1) spans most of these operators and (2) allows the implicit use of the context of that action. Thus, the instructions studied in the following describe the placement of the first brick of a new structure, i.e. giving the semantic generator the possibility to refer (or not) to the just finished one.

We describe below how we use crowdsourcing to gather human descriptions of these placements (mainly to parametrize the semantic and syntactic generator) and compare the efficiency of human vs. automatic descriptions. For this, *we asked subjects to actually perform the actions*. We expect conditions (human vs. AI-generated utterances, effective vs. no use of context) will impact placement error or time-to-complete.

All experiments are performed in French.

4.1 Collection of ground-truth data

We collected and ranked verbal descriptions performed by human subjects in two steps:

Free descriptions provide us with a ground-truth corpus of verbal descriptions of elementary placements performed by human subjects in which they put themselves in the place of a robot instructor.

Ratings of the former verbalizations were then collected by asking subjects to put themselves in the place of human partners and listen to the robot's instructions.

Scene presentation. In both sub-tasks, subjects are presented with bricks laid on a board. The brick just placed by the robot and the one to be placed by the subjects are respectively displayed with back edges vs. 50% transparency. The following *video* shows the screen during the experiment.

Instructions. The proposed vs ranked verbal descriptions should be unambiguous and as short as possible. Before the actual test (24 scenes), we

trained the participants with three examples containing some incorrect propositions. The results are used to validate crowdsourced data. We give the participants additional instructions when describing an action, namely, the use of specific terms (e.g. cube, brick, east-west orientation, north of) and the possibility to refer to the previously laid brick or any overtly constructed structure.

Subjects. The *free descriptions* were performed by the authors and 10 French-speaking participants while 15 participants recruited through Prolific performed the *ranking*.

Analysis of free descriptions. We combined descriptions performed by the authors with the ones suggested by the 10 participants. We manually selected an average of 5 *natural instructions* per scene in order to ensure that there were no duplicates, mistakes or ambiguities while trying to span as closely as possible the variety of styles, in particular syntactic constructs, topological properties of objects, etc.

Feature selection. Then, amongst all the sentences, we gather the following key parts which are essential or helpful for the action description: Hierarchy (Hierarchical Planning) and precedence between actions in the assembly (Short and long term recall when referencing objects/landmarks). The different reference types being *the previously placed element*, *a built structure* (e.g. tour, bridge, staircase etc) and *a part of an element:* (e.g. sides, corners, section of a structure etc). Aside from the reference object, an instructed action can be *decomposed* using multiple sub-actions or it can contain *verification* (i.e. additional information for validating the executed action). We note that the topological features and types of objects that can be found in our suggested scenes are taken from Borrmann and Rank (2009).

Analysis of the ranking. Following this phase, we repeat the same experiment with the same example and test sets along with the updated set of scene descriptions. However, we only ask the participants to choose the best, among the new list of natural instructions (see video). 15 participants are recruited through Prolific. The participants have to be French native speakers. The reason behind this experiment is to check which criteria a human agent would prefer as an instruction (e.g. including, or not, a verification step).

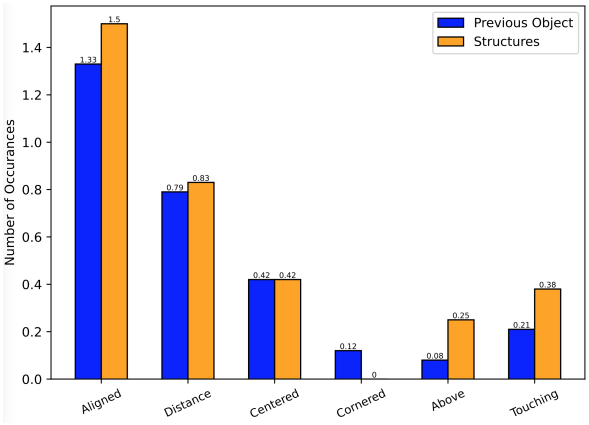

Figure 8: Frequency of appearance of topological operators in the proposed scenes

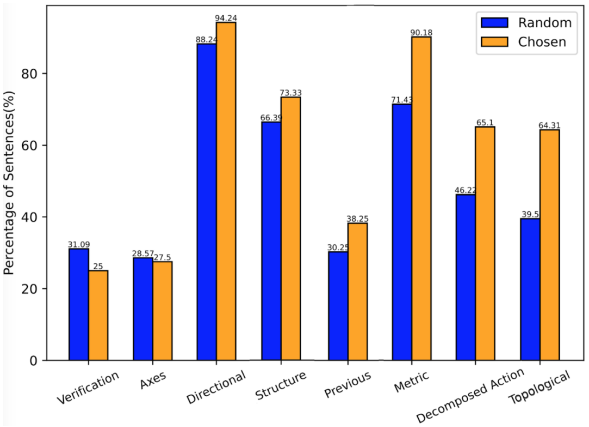

Figure 9: Results of all description criteria (by chance, and chosen by the 15 participants)

4.1.1 Results

As instructed, the participants are to choose a compact description of the action, all while being unambiguous and easy to understand. Amongst all the participants over all the scenes, 58.33% of the chosen sentences are the shortest ones. Suggesting that even with the instruction of 'choosing a compact sentence', longer sentences (usually caused by adding a verification step, or using multiple sub-actions) are also considered by the participants.

Figure 9 compares the percentage of sentences chosen at random with the ones chosen by the participants when a certain criterion is provided. In other words, the more the participants prefer a criterion, the bigger the difference will be between both percentages for that criterion. An instruction is a combination of multiple criteria, however the results still show a difference of preference between some of them. Going from the right, we see that when the feature is provided, the use of 'Decomposition' (decomposing an instruction), 'Metric' (numerical distance), and 'Topological' (touching reference)} were mainly preferred. Then 'Previous'

(mention of previously placed object), 'Structure' (mention of context and structures), and 'Directional' (3D directions) were also preferred but with a smaller percentage. The final two criteria {Verification, Axes} were disliked by the participants. Despite the fact that including a 'Verification' step ensures the correctness and reduces the ambiguity in an instruction, only 25% of the times do the participants choose the option with verification. This might be due to the fact that the participants are steering away from longer sentences containing this verification step. Lastly, we notice that the 'Axes' criterion corresponding to alignment is not largely preferred, which might be caused by the complexity of the positioning compared to other options.

Table 1: Subjective evaluation of different aspects of our verbalisation: 1:Strongly disagree - 5:Strongly agree

Questions	Score
1- Utterances were generated by a computer	4.3
2- Instructions were unambiguous	2.9
3- I prefer instructions referring to structures in place	3.98
4- Utterances were spelled clearly	3.65
5- Syntax was correct	4.15
6- I prefer instructions referring to the brick just placed by the robot	3.55
7- Utterances were generated by humans	3.33
8- The complexity of sentences were well adapted to the task	3.05

Table 2: Subjective evaluation of the participants' mental charge: 1:Strongly disagree - 5:Strongly agree

Questions	Score
1- The task was highly demanding	3.9
2- The pace of the task was too fast	2.9
3- You managed to accomplish what you were told to do	3.23
4- You worked hard to achieve your level of performance	3.9
5- You were insecure and stressed	2.48

Table 3: Different styles of sentences for the scene in fig.10

Type	Sentence
robot	Pour terminer la tour Sud, je mets un cube rouge ici.
worst_NI	Empile une barre bleue orientée Est-Ouest pour recouvrir exactement le haut de l'escalier, et le pilier qui est à l'Ouest de ce dernier.
best_NI	Dépose une barre bleue recouvrant complètement la tour rouge au Nord et le sommet de l'escalier jaune.
without_context	Place une barre bleue orientée Est-Ouest dont le côté Ouest doit être aligné avec celui du cube précédent laissant deux tenons libres vers le Nord.
with_context	Pour faire un pont, place une barre bleue orientée Est-Ouest qui recouvre le sommet jaune de l'escalier et le sommet rouge de la tour Nord.

4.2 Task Assembly

The previous experiments helped us to identify the key features used and preferred by humans for verbally instructing an action. We now test the impact of this verbal instruction on effective action performance.

Figure 10 shows an example scene and table 3 shows the sentences for that scene. The line 'Robot' gives the sentence accompanying the robot's first action. The other four sentences correspond to the

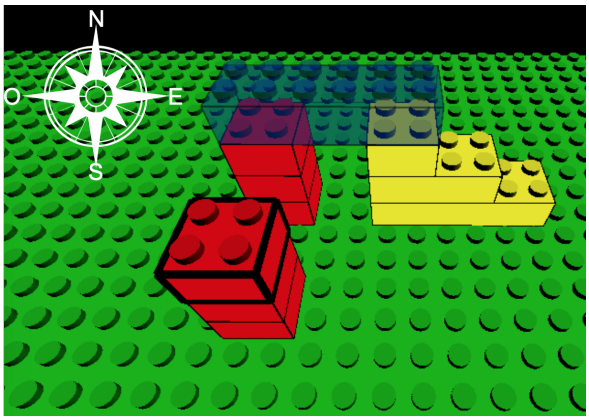

Figure 10: Scene example, having the bold red cube as the previously placed cube by the robot, and the transparent blue bar as the new object to be placed by the human agent

different styles we are comparing when instructing a participant. At first glance, we see in this example some difficulty of giving an instruction without the use of structures/context. This is why both highest and lowest ranked natural sentences use context and structures, suggesting their importance when giving the instruction.

Scene presentation. The participants were asked to observe the robot placing a brick on the game board, and then to continue the assembly according to its verbal instructions (see video). We use the same scenes as before and increase the test set using data augmentation (mirroring along the north/south axis), resulting in 54 scenes in total (3 training scenes & 51 test scenes).

Subjects. 40 French native speakers are recruited through the Prolific platform. 86.79% are right handed, 50% identified as men and 88.67% have already played with LEGO™ before this experiment.

Instructions and conditions. They have to place the right element as instructed, accurately and as fast as possible (see video). We have 4 instruction styles: (a) Lowest preference rate from the data-collection experiments ($worst\_NI$). (b) Highest preference rate from the data-collection experiments ($best\_NI$). (c) AI-generated description without mention of structures ($without\_context$). (d) AI-generated description with mention of structures ($with\_context$). The 4 styles are equally distributed among the 40 participants so that each scene with a given style is exactly performed by 10 participants.

Final questionnaires. We also include a two-part, 5-point Likert scale, questionnaire at the end of the

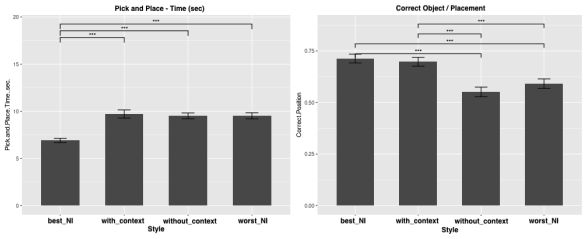

Figure 11: Normalized results over the 4 different styles – scenes with only one structure as available reference

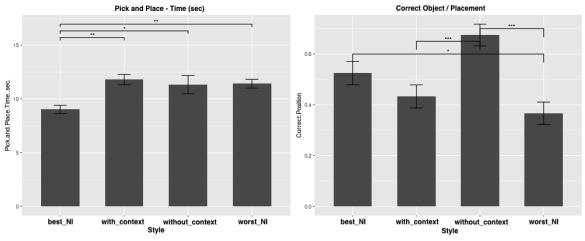

Figure 12: Normalized results over the 4 different styles – scenes with multiple structures as available references

experiment: (a) the first one evaluates the different parts of our verbalizer (table 1); (b) the second one (table 2) uses the NASA Task Load Index (NASA-TLX) (Hart and Staveland, 1988), which is the most common, subjective, multidimensional framework Colligan et al. (2015) to measure the cognitive load.

Results. For each of the 4 styles, we measure 5 cues: (a) the pick_to_place time (i.e. time between the first chosen object and the final release), (b-c) the number of chosen objects and positions to evaluate the participants' hesitation, (d-e) the percentage of correct selections and placements to evaluate performance. Results are given in Figures 11-12.

We select scenes involving more than 2 bricks but less than 2 laid structures (see Figure 11). We fit a linear mixed-effects model (*lmr* from *lme4* R package) with the scene as additional factor and subjects as random effect, and performed a post-hoc Tukey adjustments for pairwise comparisons (*ght* from *multcomp* R package). The highest ranked natural instruction significantly outperforms ($p < 10^{-3}$) the three other styles for time-to-complete and all but best-AI for successful completion. For scenes 21-24 with more than 2 laid structures (see Figure 12), AI-generated descriptions without mention of structures unexpectedly outperforms ($p < 10^{-3}$) all others for successful completion at a large margin: it seems that complex calculations seduce human intelligence but penalize performance. It also mirrors the findings of the data collection: people propose the use of axes and verifications (in experiment 1) but dislike them when asked to choose the best instruction (in experiment 2).

The verbalisation questionnaire (Tab. 1) shows that the verbalizer is working adequately: instructions are syntactically correct and clear, do not contain any major ambiguities and are properly uttered. The participants agree that both the use of hierarchical context and the mentioning of previously placed objects are a plus, while still leaning more towards the use of the former.

The NASA-TLX questionnaire (Tab. 2) shows that the experiment does require effort and cognitive load. It also shows that the participants are fairly satisfied with the rhythm, do not have much to say about their performance and do not express important signs of stress/frustration.

5 Conclusions and Future Work

We evaluate the impact of using hierarchical *context* when giving instructions in an assembly task. We gathered and ranked crowdsourced human instructions. We set up a multi-layer verbalizer that computes AI-generated instructions. We then compared the performance of these verbalization policies on a web-based assembly task. The differences between users' preferences and actual performances claim for an evaluation method in two steps: first, selecting candidate policies by subjective preference but then assess their efficiency by objective performance. We see that referring to hierarchical context improves human performance, compared to refraining from using it, in particular when context is unambiguous. We also show that our AI-generated instructions often outperform the least popular human instructions, validating the efficiency of our verbalizer.

While verbalizing, pointing towards the intended object would improve the understanding of the robot's intention, and reduce the effort in the verbal explanation to ensure task completion. Therefore, future work will include this modality in the action layer of our architecture along with speech-hand-gaze coordination. Incremental monitoring of actions by perception, in particular for on-line comprehension and attention, is a key issue for HRI.

Acknowledgments

This work is funded by MIAI (ANR-19-P3IA-0003).

References

Rachel Virgínia Xavier Aires, Aline M. P. Manfrin, Sandra M. Aluísio, and Diana Santos. 2004. What is my style? using stylistic features of portuguese web texts to classify web pages according to users' needs. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004, May 26-28, 2004, Lisbon, Portugal*. European Language Resources Association.

Rachid Alami, Raja Chatila, Sara Fleury, Malik Ghallab, and Félix Ingrand. 1998. An architecture for autonomy. *Int. J. Robotics Res.*, 17(4):315–337.

Gregor Behnke, Pascal Bercher, Matthias Kraus, Marvin R. G. Schiller, Kristof Mickeleit, Timo Häge, Michael Dorna, Michael Dambier, Dietrich Manstetten, Wolfgang Minker, Birte Glimm, and Susanne Biundo. 2020. New developments for robert - assisting novice users even better in DIY projects. In *Proceedings of the Thirtieth International Conference on Automated Planning and Scheduling, Nancy, France, October 26-30, 2020*, pages 343–347. AAAI Press.

Robert S. Belvin, Ron Burns, and Cheryl Hein. 2001. Development of the HRL route navigation dialogue system. In *Proceedings of the First International Conference on Human Language Technology Research, HLT 2001, San Diego, California, USA, March 18-21, 2001*. Morgan Kaufmann.

Dan Bohus, Chit W. Saw, and Eric Horvitz. 2014. Directions robot: in-the-wild experiences and lessons learned. In *International conference on Autonomous Agents and Multi-Agent Systems, AAMAS '14, Paris, France, May 5-9, 2014*, pages 637–644. IFAAMAS/ACM.

André Borrmann and Ernst Rank. 2009. Topological analysis of 3d building models using a spatial query language. *Adv. Eng. Informatics*, 23(4):370–385.

Gerard Canal, Senka Krivic, Paul Luff, and Andrew Coles. 2021. Task plan verbalizations with causal justifications. In *ICAPS 2021 Workshop on Explainable AI Planning (XAIP)*.

Justine Cassell and Timothy Bickmore. 2003. Negotiated collusion: Modeling social language and its relationship effects in intelligent agents. *User modeling and user-adapted interaction*, 13(1):89–132.

L. Colligan, H. Potts, Chelsea T. Finn, and R. A. Sinkin. 2015. Cognitive workload changes for nurses transitioning from a legacy system with paper documentation to a commercial electronic health record. *International journal of medical informatics*, 84 7:469–76.

Fethiye Irmak Dogan, Sarah Gillet, Elizabeth J. Carter, and Iolanda Leite. 2020. The impact of adding perspective-taking to spatial referencing during human-robot interaction. *Robotics Auton. Syst.*, 134:103654.

Michelangelo Fiore, Aurélie Clodic, and Rachid Alami. 2014. On planning and task achievement modalities for human-robot collaboration. In *Experimental Robotics - The 14th International Symposium on Experimental Robotics, ISER 2014, June 15-18, 2014, Marrakech and Essaouira, Morocco*, volume 109 of *Springer Tracts in Advanced Robotics*, pages 293–306. Springer.

Kate Forbes-Riley, Diane Litman, and Mihai Rotaru. 2008. Responding to student uncertainty during computer tutoring: An experimental evaluation. In *International Conference on Intelligent Tutoring Systems*, pages 60–69. Springer.

Rachel Gockley, Allison Bruce, Jodi Forlizzi, Marek P. Michalowski, Anne Mundell, Stephanie Rosenthal, Brennan Sellner, Reid G. Simmons, Kevin Snipes, Alan C. Schultz, and Jue Wang. 2005. Designing robots for long-term social interaction. In *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems, Edmonton, Alberta, Canada, August 2-6, 2005*, pages 1338–1343. IEEE.

Guillermo Gomez, Carole Plasson, Frédéric Elisei, Frédéric Noël, and Gérard Bailly. 2015. Qualitative assessment of an immersive teleoperation environment for collaborative professional activities in a" beaming" experiment. In *EuroVR 2015-European conference for Virtual Reality and Augmented Reality*, pages 8–pages.

Sandra G Hart and Lowell E Staveland. 1988. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology*, volume 52, pages 139–183. Elsevier.

Jussi Karlgren. 2000. *Stylistic Experiments for Information Retrieval*. Ph.D. thesis, Royal Institute of Technology, Stockholm, Sweden.

Hiroaki Kitano, Minoru Asada, Yasuo Kuniyoshi, Itsuki Noda, and Eiichi Osawa. 1997. Robocup: The robot world cup initiative. In *Proceedings of the First International Conference on Autonomous Agents, AGENTS 1997, Marina del Rey, California, USA, February 5-8, 1997*, pages 340–347. ACM.

David Kortenkamp, Reid Simmons, and Davide Brugali. 2016. *Robotic Systems Architectures and Programming*, pages 283–306. Springer International Publishing, Cham.

Alastair MacFadden, Lorin Elias, and Deborah Saucier. 2003. Males and females scan maps similarly, but give directions differently. *Brain and Cognition*, 53(2):297–300.

François Mairesse and Marilyn A. Walker. 2010. Towards personality-based user adaptation: psychologically informed stylistic language generation. *User Model. User Adapt. Interact.*, 20(3):227–278.

Juliana Miehle, Wolfgang Minker, and Stefan Ultes. 2018a. What causes the differences in communication styles? A multicultural study on directness and

elaborateness. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*. European Language Resources Association (ELRA).

Juliana Miehle, Wolfgang Minker, and Stefan Ultes. 2018b. What causes the differences in communication styles? a multicultural study on directness and elaborateness. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Paul Molins and Guy Lapalme. 2015. Jsrealb: A bilingual text realizer for web programming. In *ENLG 2015 - Proceedings of the 15th European Workshop on Natural Language Generation, 10-11 September 2015, University of Brighton, Brighton, UK*, pages 109–111. The Association for Computer Linguistics.

Stefanos Nikolaidis, Minae Kwon, Jodi Forlizzi, and Siddhartha S. Srinivasa. 2017. Planning with verbal communication for human-robot collaboration. *CoRR*, abs/1706.04694.

Damien Pellier and Humbert Fiorino. 2018. PDDL4J: a planning domain description library for java. *J. Exp. Theor. Artif. Intell.*, 30(1):143–176.

Vittorio Perera, Sai P. Selvaraj, Stephanie Rosenthal, and Manuela M. Veloso. 2016. Dynamic generation and refinement of robot verbalization. In *25th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN 2016, New York, NY, USA, August 26-31, 2016*, pages 212–218. IEEE.

Ronald Petrick and Mary Ellen Foster. 2013. Planning for social interaction in a robot bartender domain. In *Proceedings of the International Conference on Automated Planning and Scheduling*, volume 23, pages 389–397.

David Reitter, Frank Keller, and Johanna D. Moore. 2006. Computational modelling of structural priming in dialogue. In *Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 4-9, 2006, New York, New York, USA*. The Association for Computational Linguistics.

Stephanie Rosenthal, Sai P Selvaraj, and Manuela M Veloso. 2016. Verbalization: Narration of autonomous robot experience. In *IJCAI*, volume 16, pages 862–868.

Svetlana Stenchikova and Amanda Stent. 2007. Measuring adaptation between dialogs. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue, SIGdial 2007, Antwerp, Belgium, September 1-2, 2007*, pages 166–173. Association for Computational Linguistics.

Manuela M. Veloso, Nicholas Armstrong-Crews, Sonia Chernova, Elisabeth Crawford, Colin McMillen, Maayan Roth, Douglas L. Vail, and Stefan Zickler. 2008. A team of humanoid game commentators. *Int. J. Humanoid Robotics*, 5(3):457–480.

Dirk Voelz, Elisabeth André, Gerd Herzog, and Thomas Rist. 1998. Rocco: A robocup soccer commentator system. In *RoboCup-98: Robot Soccer World Cup II*, volume 1604 of *Lecture Notes in Computer Science*, pages 50–60. Springer.

Terry Winograd. 1972. Shrdlu: A system for dialog.

Terry Winograd. 1974. Five lectures on artificial intelligence. Technical report, STANFORD UNIV CA DEPT OF COMPUTER SCIENCE.

Qingxiaoyang Zhu, Vittorio Perera, Mirko Wächter, Tamim Asfour, and Manuela M. Veloso. 2017. Autonomous narration of humanoid robot kitchen task experience. In *17th IEEE-RAS International Conference on Humanoid Robotics, Humanoids 2017, Birmingham, United Kingdom, November 15-17, 2017*, pages 390–397. IEEE.