Research Article

Avinash Kumar Singh, Neha Baranwal, Kai-Florian Richter*, Thomas Hellström, and Suna Bensch

Verbal explanations by collaborating robot teams

https://doi.org/10.1515/pjbr-2021-0001
received January 31, 2020; accepted June 3, 2020

Abstract: In this article, we present work on collaborating robot teams that use verbal explanations of their actions and intentions in order to be more understandable to the human. For this, we introduce a mechanism that determines what information the robots should verbalize in accordance with Grice's maxim of quantity, i.e., convey as much information as is required and no more or less. Our setup is a robot team collaborating to achieve a common goal while explaining in natural language what they are currently doing and what they intend to do. The proposed approach is implemented on three Pepper robots moving objects on a table. It is evaluated by human subjects answering a range of questions about the robots' explanations, which are generated using either our proposed approach or two further approaches implemented for evaluation purposes. Overall, we find that our proposed approach leads to the most understanding of what the robots are doing. In addition, we further propose a method for incorporating policies driving the distribution of tasks among the robots, which may further support understandability.

Keywords: understandable robots, robot teams, explainable AI, human-robot interaction, natural language generation, Grice's maxim of quantity, informativeness

1 Introduction

Robots are becoming increasingly autonomous and capable, which entail that interacting humans need to understand the actions, intentions, capabilities, and limitations of these autonomous robots. The term *understandable robots* [1] refers to the robot's ability to make itself understood – implicitly or intentionally and in a way that interacting humans understand. In other words, *understandable* refers to others' ability to make sense of, explain, and predict a robot's behavior. Generally speaking, robots are not understandable if they act without considering the effects of their actions on the interacting human. For example, a robot that does not act on a human's request (due to whatever reason) is not understandable, and so is a robot that performs actions that are not interpretable by a human. Understandable robots increase the interaction quality [2] for humans in terms of both user experience and perception of being safe around these robots at all times.

In human-robot interaction (HRI), the number of interacting robots and humans may vary. Most work on HRI concerns interaction between one robot and one human only, even though more complex configurations of humans and robots are possible [3]. For example, a guiding robot in a museum may interact with several humans at the same time. Research on teams of several robots interacting with one or several humans is still rare, even if such configurations may very well become common in future real-world applications. For example, the covid-19 pandemic makes it more likely that groups of robots will be used to monitor body temperatures and symptoms of visitors in hospitals.

In many of these HRI configurations, humans may take different roles in the interaction with the robots, such as operator, supervisor, user, or information consumer [4]; and each role poses specific challenges to the interaction design.

In this article, we describe work that addresses understandable robot teams, where several robots collaborate with each other and explain their actions and intentions in natural language, so that human bystanders can understand what the robots are doing. The term *human bystanders* refers to humans who are concerned with, but not directly involved in, what the robots are doing (e.g., supervisors, people who need to make sure that the robots' actions pose no threat to them, or humans in close proximity to these robot teams).

* **Corresponding author: Kai-Florian Richter,** Department of Computing Science, Umeå University, Sweden, e-mail: kaifr@cs.umu.se
Avinash Kumar Singh, Neha Baranwal, Thomas Hellström, Suna Bensch: Department of Computing Science, Umeå University, Sweden, e-mail: avinash@cs.umu.se, neha@cs.umu.se, thomash@cs.umu.se, suna@cs.umu.se

Earlier work on verbal communication between robots and humans most often concerns humans giving verbal commands to robots [5]. However, also systems with robots that speak have been developed and studied. As noted in [6], most such work employs relatively simplistic processes for generating the robots' utterances. Still, in [7] a user study confirms that both user experience and performance increase if a robot gives verbal feedback on its own and others' actions to collaborating humans. The authors of [8] investigate a robot that gives both verbal commands and state-conveying actions to its human teammate, for example, "Let's rotate the table clockwise" and "I think I know the best way of doing he task," respectively. In human subject experiments, they show for their scenario that commands, rather than state-conveying actions, is the most effective method to retain user trust in the robot. Note, though, that in our article we are investigating human understanding of the robot (not trust in the robot). Generation of verbal explanations for robot behaviors was also investigated in [9] and [10], with the aim of making the robots' behavior understandable to nonexpert human users.

The current article addresses two aspects of understandable collaborating robot teams. First, it builds on our work on an architecture for collaborating robot teams [11], where the robots jointly execute a plan and also verbally comment on their current and immediate future actions. The current article outlines how the execution of actions may be governed by *policies*, for example, to evenly distribute work among all robots. Second, and more importantly, we propose how Grice's principle of informativeness [12] can be maintained by a team of robots that collaborates to solve a task while explaining ongoing and planned actions to human bystanders. More specifically, we propose an algorithm to identify sequences of actions that may be subject to verbal descriptions that respect this principle of informativeness. The algorithm decides on the most informative content that should be verbalized, where the content is a piece of information. This content is verbalized using predefined templates.

To the best of our knowledge, informative verbal explanations have not previously been addressed for talking robot teams. We hypothesize that following such a mechanism in generating verbal explanations will lead to a better understanding in human bystanders of what the robots are doing. To investigate this hypothesis, we evaluated the proposed mechanism in an empirical study, in which human participants answered a range of questions about the robots' explanations generated by either the proposed algorithm or one of two other mechanisms implemented for evaluation purposes.

The article is organized as follows. Section 2 introduces the previously developed framework for collaborative planning and execution of actions. Section 3 describes the policy approach, followed by our approach for generation of verbal utterances in Section 4. Section 5 presents and discusses the results of the empirical evaluation. Section 6 concludes the article and discusses future work.

2 A collaborating robot team

For brevity, the basic mechanisms of plan derivation and plan execution are described by referring to an example with a team of three robots before we introduce our approach to using policies for task distribution in the following section. For detailed descriptions of plan derivation and execution, see [11].

Our example consists of a team of three robot agents A, B, and C that collaborate on planning and solving a given task, which comprises moving objects R, G, and Y to a goal configuration in a 3×3 grid with cells numbered 1,2,3,4,5,6,7,8,9 (Figures 1 and 2). Each robot can only reach a limited number of cells. Thus, collaboration is necessary to reach the goal. First, the robots collaboratively derive a plan by taking into account the initial configuration on the tabletop and their individual capabilities to perform actions [11]. The plan represents the shortest sequence of robot actions to reach a given goal following a specific policy (see Section 3). During collaborative plan execution, the robots utter natural language sentences that explain what they will do or what they request other robots to do.

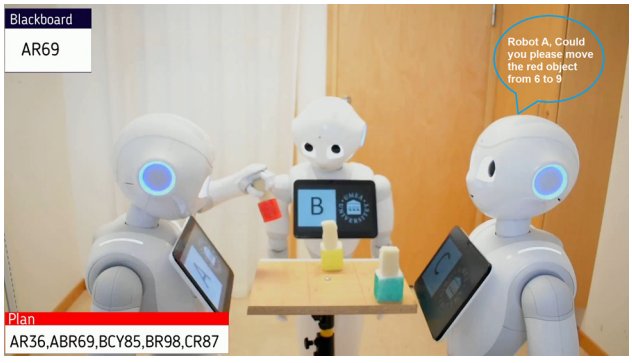

Figure 1: Experimental setup of collaborating Pepper robots moving objects while explaining their own and others' actions.

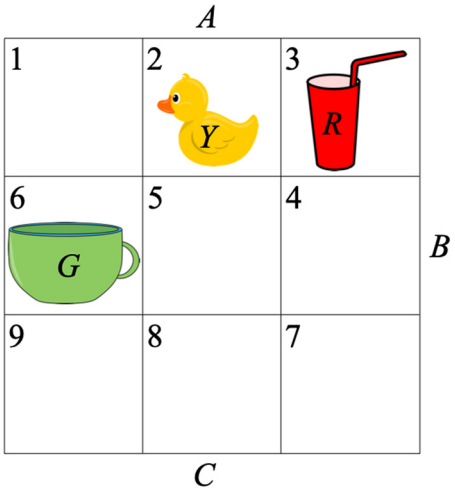

Figure 2: A numbered grid with three objects: a red glass (R), a green cup (G), and a yellow duck (Y). The objects are placed in different initial cells of the grid. Three robots A, B, and C are placed around the tabletop, and each robot can reach only some of the cells.

Such a tabletop setting presents a rather simple and somewhat artificial scenario. But this scenario was chosen exactly because of its simplicity, and because it, therefore, offers substantial experimental control and proof-of-concept possibility of the developed methods and algorithms. It also allows for focusing on "high-level" aspects, such as task distribution or verbalization strategies, without a large "overhead" resulting from complicated robot operations or complex environments. Our findings may then be later transferred to more realistic scenarios, where the understandability of robot teams becomes more relevant, for example, in search and rescue scenarios, in exploration (or cleaning) of hazardous sites, or simply scenarios in industry or household, where human supervisors need to be sure that robots indeed execute the intended plan.

Coming back to the example scenario outlined above, for a goal of moving object R to cell 6, and an initial grid as illustrated in Figure 2, the following plan s is assumed to have been derived:

$$ABR34, ABCR45, AG61, ACR56.$$

Plan s is a string composed of a sequence of *action templates* a_i separated by commas: $s = a_1, a_2, \ldots, a_N$. For example, the substring $ACR56$ is an action template, denoting that either robot A or C would be capable of moving object R from cell 5 to cell 6. Hence, an action template may be underspecified in the sense that it allows different robots to execute the same action. Deciding on a specific robot to execute the action is denoted as *instantiation* of the action template, and the result is an *action* with the same form as an action template but with only one robot identifier. For example, instantiation of the action template $a_1 = ABR34$ may result in the action $\alpha_1 = BR34$.

3 Policies for instantiating action templates

For most plans, there will be different possible instantiations of a given plan. Arguably, some of these may be smarter than others. For example, in a sequence of action templates, it may be best if a single robot executes as many actions in a row as possible in order to avoid overhead in task switching or maybe even to allow skipping intermediate steps. Given plan s above, in principle robot A could move object R directly from cell 4 to 6 without first going through cell 5. Alternatively, it might be best to avoid having a robot execute more than one action in a row, for example, to allow for cooldown between action execution. In this case, an optimal sequence of actions α_i for plan s may be $BR34, CR45, AG61, CR56$.

More generally, each of the action templates in s may contain up to three different robots A, B, C that may execute an action. The task is to find for each action template a_i the best instantiation, i.e., the optimal action α_i. However, how to best instantiate an action template a_i may depend on the choices made for a previous template a_{i-1} or a coming template a_{i+1}. We use the term *policy* to denote the principle that governs the instantiations of action templates.

A penalty is assigned for each instantiation that violates the constraints imposed by a given policy. The sequence of actions with the lowest overall penalty is the optimal one according to the policy. In our example, the number of possible instantiations is very limited. There are four action templates, and each contains at most three different robots, i.e., there are at most $3^4 = 81$ possibilities and in practice far less. Thus, we can simply calculate the penalty for every possible sequence of instantiations and then pick the one with the lowest total penalty. If the number of action templates and/or the number of robots increases, more efficient algorithms may be used that use look-ahead and backtracking mechanisms to find an optimal instantiation (cf. [13]).

In the general case, for a given policy p, the optimal instantiation of a sequence of action templates $a_1, a_2, \ldots, a_M$ is the solution to the following minimization problem:

$$(\alpha_1, \alpha_2, \ldots, \alpha_M) = \underset{I \in I_1 X I_2 X \ldots X I_M}{\arg\min} \text{penalty}_p(I), \qquad (1)$$

where I_i is the set of all possible instantiations of action template a_i, and $\text{penalty}_p(I)$ is the penalty to perform action sequence I under policy p. For plan s, $I_1 = \{AR34, BR34\}$, $I_2 = \{AR45, BR45, CR45\}$, $I_3 = \{AG61\}$, $I_4 = \{AR56, CR56\}$; and the optimal $I_o = (BR34, CR45, AG61, CR56)$ given a policy p of uniform workload.

3.1 Potential policies

A range of policies may be implemented. Every robot may use the same policy, or each may use a different one. Both the decisions on what policy to use and whether or not all robots employ the same policy may depend on different factors, for example, the kind of robots involved, the tasks at hand, or the environment the robots interact in. Our approach is flexible in that respect.

- **Random**: such a (non-)policy randomly picks one of the robots listed in an action template. There are no needs or preferences to prioritize one way of executing actions over any other and, accordingly, no penalties are applied.
- **Least actor changes**: aims at minimizing the number of times action execution switches between robots. In other words, the policy aims at maximizing the sequence of actions performed by a single robot. A penalty of 1 is assigned each time the action execution switches from one robot to another.
- **No repeat actors**: this is essentially the opposite of the previous policy. Each action should get assigned to a robot different from the one that executed the previous action. This may be useful if executing an action is strenuous on the robots or resetting to a pose allowing to execute the next action takes a long time. A penalty of 1 is applied each time a switch to a new robot is not possible.
- **Highest variation**: similar in effect, but different in intention, this policy aims at including as many different robots as possible in action execution to increase variability. A penalty of 1 is applied each time a robot that has already been selected to execute an action gets selected again.
- **Uniform workload**: aims at engaging all robots to the same extent. A penalty would be assigned to an entire sequence of actions to quantify how much, or little, all robots are engaged, compared to the average robot. The entropy of the action distribution may be a useful measure for the penalty in this case.
- **Lazy robot**: the robot that instantiates the action templates avoids to perform actions itself. It picks other robots to do a task anytime this is possible. This policy operates locally only on the robot equipped with it. All other robots might still assign tasks to it without a penalty.
- **Avoid robot**: turning the lazy robot policy into a global policy, where every robot aims not to assign actions to a particular robot, incurring a penalty of 1 if that robot gets to execute an action. Such a policy may be useful if one of the robots runs out of battery or loses some capability to execute actions, for example.

Note that there may be several optimal instantiations for a given plan and policy. In addition, the policies are not mutually exclusive; different policies may lead to identical instantiations. There are other possible policies, including applying combinations of policies. For example, combining the uniform workload and avoiding robot policies would aim at an even distribution of workload among all involved robots except for one designated robot that should not do any work.

4 Verbalization of actions

In the following, we describe an algorithm that divides an instantiated plan into partitions that the robots may describe verbally. The partitions are created to support Grice's maxim of quantity or informativeness [12]: that a speaker should be as informative as possible, while at the same time not say more than is required. Generating utterances this way is motivated by the assumption that such explanations will lead to a good understanding in human bystanders of what the robots are doing. The algorithm decides on the most informative content that should be verbalized, where the content is an abstract piece of information representing robot actors and specific objects and from where to where the objects are moved on the tabletop. The content is verbalized using predefined templates that specify sequences and sets of possible words and phrases that a robot can choose from. Thus, currently we do not utilize the traditional pipeline of document planning, microplanning, and surface realization [14].

We assume that the entire plan has been instantiated according to a chosen policy. This assumption mostly allows for an easier explanation of the proposed algorithm in the following. It is not a principle restriction of the presented approach. In fact, equation (1) allows for incremental and flexible execution of robot actions. For example, assume that each time only two subsequent

action templates get instantiated and immediately executed by the robots. Then, in principle, the instantiation of the next two subsequent action templates could be done according to how the execution of the previous two actually went, which among others would allow for adapting the policy used in their instantiation.

However, for simplicity reasons we assume an instantiation of the entire plan according to a chosen policy. This corresponds to solving the minimization problem in equation (1) for all action templates, i.e., $M = N$, where N is the length of a given plan s. For understandability, the relevant information is what is being said (verbalized referents or knowledge in a given discourse) and who is saying it. We distinguish between newly introduced information and information that does not change for the next immediate planned action. For example, if a robot says, "*I will move object G*" and another robot then says "*I can move it further*," then object G is nonchanging information; whereas the second robot speaking is newly introduced information, i.e., a speaker change. Newly introduced information should be made explicit, whereas nonchanging information can be referred to or go unmentioned. This aligns with the above maxim that an utterance should be as informative as possible (explicitly mentioning newly introduced referents or speaker change), while at the same time being concise (assuming nonchanging information as known and not explicitly verbalizing it).

Given an instantiation $\alpha_1, \alpha_2, \ldots, \alpha_N$, each action α_i consists of four *information units* $r_i o_i f_i t_i$, denoting a robot, object, source, and destination. For example, the action $CR56$ consists of $r_i = C$, $o_i = R$, $f_i = 5$ and $t_i = 6$. For a given instantiation, a *partition* is any substring of successively concatenated α_i, where either the robot r_i or the object o_i does not change. The maximum length of a partition is set to 4 in order not to overload a human bystander with too much information at once.

Algorithm 1 returns the set of all partitions for a given instantiation. The algorithm is similar to the Cocke–Younger–Kasami algorithm [15]. It checks for successive substrings α whether the acting robot or the object to be moved changes. Let, for example, $AY23, BY36, CG41, AG12, AG25$ be an instantiation. Here, $\alpha_1, \alpha_2 = AY23, BY36$ is a partition since the object Y is the same in both α_1 and α_2. Whereas $\alpha_2, \alpha_3 = BY36, CG41$ is not a partition since the robot and the object is different in α_2 and in α_3. Similarly, $\alpha_1, \alpha_2, \alpha_3 = AY23, BY36, CG41$ is not a partition since neither the robots nor the objects are the same for all $\alpha_1, \alpha_2, \alpha_3$. Once the algorithm computes the set of all partitions for a given instantiation, we pick the partitioning that covers the entire plan, consists of nonoverlapping successive partitions, and has the least changes in information units as the *optimal sequence*. That is, the optimal sequence is a sequence of successive, nonoverlapping partitions that represent an entire sequence of actions where either the robots or the objects are the same.

The verbalization of the optimal sequence supports avoiding underinformative and overinformative utterances since newly introduced information is explicitly verbalized, whereas nonchanging information is not. We now give the algorithm and exemplify it later with the instantiation example mentioned above.

Algorithm 1

Input: an instantiation $\alpha_1, \alpha_2, \ldots, \alpha_N$, where each α_i has four information units $r_i o_i f_i t_i$.

Output: the set of all partitions ϕ of the input $\alpha_1, \alpha_2, \ldots, \alpha_N$.

Method:

1. Let each α_i be a partition ϕ_i, $1 \le i \le N$.
2. If for two consecutive actions $\alpha_i \alpha_{i+1}$, $1 \le i \le N-1$ the following holds:

$$r_i = r_{i+1} \text{ or } o_i = o_{i+1}$$

   let $\phi_{ii+1} = \alpha_i \alpha_{i+1}$ be a partition.
3. If for three consecutive actions $\alpha_i \alpha_{i+1} \alpha_{i+2}$, $1 \le i \le N-2$ the following conditions hold:

   (a)

$$r_i = r_{i+1} = r_{i+2} \text{ or } o_i = o_{i+1} = o_{i+2}$$

   let $\phi_{ii+1i+2} = \alpha_i \alpha_{i+1} \alpha_{i+2}$ be a partition;

   (b)

$$r_i = r_{i+1} \text{ or } o_i = o_{i+1}$$

   let ϕ_{ii+1}, ϕ_{i+2} be the two partitions for $\alpha_i \alpha_{i+1}$ and α_{i+2};

   (c)

$$r_{i+1} = r_{i+2} \text{ or, } o_{i+1} = o_{i+2},$$

   let ϕ_i, ϕ_{i+1i+2} be the two partitions for α_i and $\alpha_{i+1} \alpha_{i+2}$.
4. If for four consecutive actions $\alpha_i \alpha_{i+1} \alpha_{i+2} \alpha_{i+3}$, $1 \le i \le N-3$ the following conditions hold:

   (a)

$$r_i = r_{i+1} = r_{i+2} = r_{i+3} \text{ or } o_i = o_{i+1} = o_{i+2} = o_{i+3}$$

   let $\phi_{ii+1i+2i+3} = \alpha_i \alpha_{i+1} \alpha_{i+2} \alpha_{i+3}$ be one partition;

   (b)

$$r_i = r_{i+1} = r_{i+2} \text{ or } o_i = o_{i+1} = o_{i+2}$$

   let $\phi_{ii+1i+2}, \phi_{i+3}$ be two partitions for $\alpha_i \alpha_{i+1} \alpha_{i+2}$ and α_{i+3};

(c)

$$r_{i+1} = r_{i+2} = r_{i+3} \text{ or } o_{i+1} = o_{i+2} = o_{i+3}$$

let $\phi_i, \phi_{i+1i+2i+3}$ be two partitions for α_i and $\alpha_{i+1}\alpha_{i+2}\alpha_{i+3}$;

(d)

$$r_i = r_{i+1} \text{ or } o_i = o_{i+1}$$

and

$$r_{i+2} = r_{i+3} \text{ or } o_{i+2} = o_{i+3}$$

let $\phi_{ii+1}, \phi_{i+2i+3}$ be two partitions for $\alpha_i\alpha_{i+1}$ and $\alpha_{i+2}\alpha_{i+3}$.

Let us again consider the example instantiation *AY*23, *BY*36, *CG*41, *AG*12, *AG*25. Figure 3 shows the set of all possible partitions. The smallest number of non-overlapping partitions that cover the entire instantiation constitutes an *optimal sequence of partitions*. Finding the optimal sequence can be implemented as a greedy search or any other suitable approach that sorts and combines the partitions into the optimal sequence. In our example, the optimal sequence of partitions is $\phi_{12}\phi_{345}$. The robots may use an optimal sequence of partitions to generate natural language utterances that support Grice's maxim of quantity. For example, for $\phi_{12} = AY23BY36$, robot *A* could utter "*I can move the yellow object from 2 to 3 and then robot B can move it from 3 to 6*," robot *B* could say "*Robot A, move the yellow object from 2 to 3, then I can move it from 3 to 6*," or robot *C* could say "*Robot A, please move the yellow object from 2 to 3, then B can move it from 3 to 6*."

A more flexible verbalization strategy making it possible to verbalize only the chosen information units (and not all) could lead to utterances such as "*I can move the yellow object towards robot B*," "*Robot A, move object Y from cell 2 to 3, and then B can move it further to cell 6*," or "*I suggest robots A and B move Y from cell 2 to 6*." Verbalizing only specifically chosen information units (e.g., mentioning objects but not all cell numbers) would most likely impact the conveyed levels of information. In our experiments, we chose to always verbalize all involved robots, cell numbers, and objects, in order to have a uniform base for comparison and evaluation as discussed in the following section.

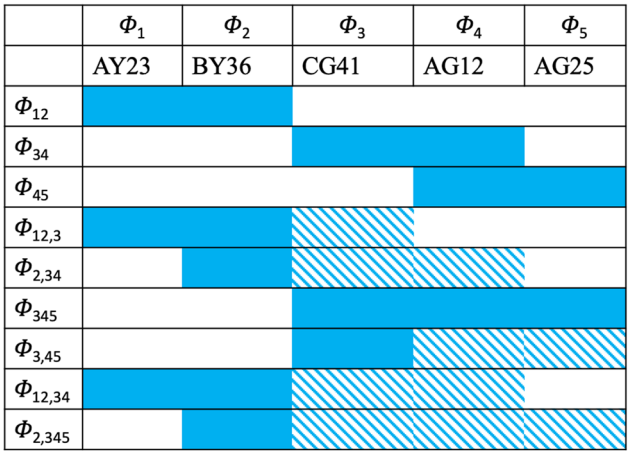

Figure 3: A visualization of all possible partitions ϕ for the instantiation *AY*23, *BY*36, *CG*41, *AG*12, and *AG*25. A partition is indicated by successive columns in solid or striped blue.

5 Empirical evaluation

We evaluated the proposed approach to generate verbal explanations, and in particular the policy of following Grice's maxim of informativeness in these explanations, in an empirical evaluation, in which human participants, who were naive with respect to our work, answered a range of questions about the robots' explanations.

5.1 Methods

We implemented three different mechanisms for generating verbal explanations on the Pepper robots. The first (called "optimal" in the following) follows Grice's maxim of informativeness [12] and is implemented as explained in Section 4.

The second mechanism forces robots to verbalize each action separately. For every action of the plan, one of the robots is identified to provide a verbal explanation, following the principles explained in Section 2. In other words, an action gets verbalized and then executed, then the next action gets verbalized and executed, and so on. We term this mechanism "single."

A third mechanism randomly segments actions to be verbalized. Accordingly, we call it "random." The robot that gets to verbalize the next sequence of actions picks between one to four of the next action templates in the plan, instantiates them, and then verbalizes this part of the plan.

For the evaluation, we fixed the plan to the following instantiated action templates. We used a fixed plan to ensure that any differences in robot behavior would only result from the different segmentation/verbalization policies.

$$AR12, AR23, BY45, BR34, BR47, CR78, CR89 \quad (2)$$

That is, robot A is to move the red object from cell 1 to cell 3, then robot B moves the yellow object "out of the

way" from cell 4 to cell 5, followed by it moving the red object from cell 3 to cell 7. Finally, robot C moves the red object from cell 7 to cell 9. This plan instantiation follows the "uniform workload" policy outlined in Section 3, and incidentally also the "least actor changes" one.

We recorded several videos of the robots performing this plan using either of the three different utterance mechanisms. From those, we selected one video each and created three variants of an otherwise identical online questionnaire.[1] In the first variant, questionnaire A, the participants got to see the video following the "optimal" mechanism first, followed by the one using the "single" mechanism, and finally the video with the "random" mechanism. Variant B showed the "random" mechanism video first, then the "optimal" mechanism one, then the "single" mechanism video. Variant C had "single" first, then "random," and then "optimal."

In each variant, participants would watch a video, followed by a range of questions asking them how they perceived the robots' explanations on a scale of 1 to 5 (an "intrinsic" evaluation according to [16]). Each of these scales would represent an attribute pair following the Godspeed questionnaire [17]; see Table 1 for all attribute pairs. After answering these questions, participants continued to the next video. However, after the first video, we first asked them to write down the plan the robots executed, which would test for their understanding and memory of what had just happened (a performance evaluation or an "extrinsic" one according to [16]). After all three video/question sets, we finally asked participants about their general preferences by having them rank the three videos from "like best" via "in-between" to "like least" (an intrinsic evaluation again). Before submitting their answers, participants also had the chance to leave some comments, which was optional.

Overall, eight participants replied to the questionnaire variants *A* and *B* each, and nine participants to variant *C*.

We scored the plans, i.e., the descriptions of what the robots had been doing, which the participants reproduced from memory using the following set of rules:

- Score starts at 0.
- For every correct move (e.g., a move from cell 1 to cell 2), add 1 point.
  - Not specifying intermediate steps was taken to indicate a correct sequence of moves and would give full points; e.g., "move from 1 to 3" is seen to implicitly contain cell 2 as an intermediate step. Hence, this would give 2 points.
- For every correct assignment of a robot to a move (e.g., "A moves from 1 to 2"), add 1 point.
  - Following the above rule, "*A* moves from 1 to 3" would give 2 points.
- For every correct object assignment to a move (e.g., "R is moved from 1 to 2"), add 1 point.
  - Again, "*R* is moved from 1 to 3" would give 2 points.
- For any sequence violation, subtract 2 points.
  - For example, "AR13,BR37,BY45" are all correct moves according to the plan, but the second and third steps are in the wrong order. Thus, 2 points would be deduced here.
- The final score is the maximum of 0 and the accumulated points.

Consequently, the minimum points a participant can achieve are 0, the maximum 21. These points are then normalized simply by dividing them by 21. The points participants achieve reflect their understanding and memory of what the robots were doing; we term this *memory score* in the following.

Table 1: The different attribute pairs used in assessing participants' perception of how the robots explained and instructed each other. On a 1 to 5 Likert scale, a ranking of 1 corresponds to "totally agree" to the left most attribute of the pair, a ranking of 5 to "totally agree" with the right most attribute. The order of presenting these pairs was shuffled for each video and each participant, and they appeared in random order

1	5
Fake	Natural
Machinelike	Humanlike
Artificial	Lifelike
Static	Interactive
Inefficient	Efficient
Unpleasant	Pleasant
Incompetent	Competent
Unintelligent	Intelligent

5.2 Results

In the analysis of the questionnaire data, we are particularly interested in the differences in the ratings of the different videos, i.e., which explanations by the robots

[1] The videos can be found here. For the "optimal" condition: https://play.umu.se/media/t/0_58k7fgma; for the "single" condition: https://play.umu.se/media/t/0_rlt6nki9; and for the "random" condition: https://play.umu.se/media/t/0_yglw697e.

participants prefer. Furthermore, differences in how well participants were able to remember the plan executed by the robots are of interest, i.e., differences in their performance. Tables 2 and 3 show summary statistics for these two variables (remember that memory scores are normalized between 0 and 1). Videos were ranked by all 25 participants, so the statistics in Table 2 are based on 25 answers. But each participant only performed the memory test for the first video they saw. Accordingly, in Table 3 statistics are based on 8 (for "optimal" and "random" mechanism, respectively) and 9 (for the "single" mechanism) answers.

In addition, we asked participants how they perceived the robots' explanations along different scales and, accordingly, tested for any differences in these ratings between the three different videos (again, these ratings were done by all 25 participants). Finally, ratings of these individual scales might determine (to some extent) how participants liked the way the robots explained their actions, i.e., correlate with the overall ranking. Table 4 shows summary statistics for these questions.

The ranking (see Table 2) seems to indicate a clear preference for the "random" mechanism video, followed by the "optimal" mechanism one, and then the "single." Indeed, a significant statistical difference was observed between them ($\chi^2 = 14.682$, $df = 2$, $p = 0.001$).

The memory scores (see Equation 2 for the plan to reproduce) seem to show differences as well. Here, on average participants scored highest for the "optimal" mechanism video, closely followed by scores for the "single" mechanism video. Average score for the "random" mechanism video is less than half of that for the "optimal" one. Still no significant difference ($\chi^2 = 3.123$, $df = 2$, $p = 0.21$) was observed, likely because of a large variety in the memory scores as can be seen in the box plot of the achieved scores shown in Figure 4 (e.g., for the "random" mechanism video, the standard deviation is higher than the average score).

No statistical differences were observed for any of the behavior questions between the three videos "optimal" mechanism, "random" mechanism, and "single" mechanism. Looking at the median values in Table 4, many indicate a "neutral" rank (3 of 5), with a tendency to perceive behavior to be rather "machinelike" than "humanlike" and "artificial" than "lifelike" (median of 2 for several conditions), but also the robots to rather behave "efficient," "pleasant," and "competent" (median of 4 each). We can also observe that the mean for all but the second and third questions is highest for the "random" mechanism video, i.e., in this condition participants tended to perceive the robots on average to be (at least slightly) more "natural," "interactive," "efficient," "pleasant," "competent," and "intelligent."

A Spearman's rank correlation analysis shows several significant correlations between the different questions for each video. However, no statistically significant correlations were observed between these questions and the preference ranking for the "optimal" mechanism video. For the "random" mechanism video, statistically significant moderate to strong negative correlations were observed between questions "static–interactive" ($\rho = -0.61$, $p = 0.001$) and "inefficient–efficient" ($\rho = -0.65$, $p = 0.000$) and the preference ranking. For the "single" mechanism video, a statistically significant moderate negative correlation was observed between question "artificial–lifelike" ($\rho = -0.42$, $p = 0.036$) and the preference ranking.[2]

Table 2: Summary statistics for ranking the different videos

Statistic	Best rank	Worst rank	Mean	St. Dev.	Median
Optimal	1	3	2.080	0.702	2
Random	1	3	1.520	0.823	1
Single	1	3	2.400	0.707	3

Table 3: Summary statistics of the achieved memory scores for reproducing the plan using a normalized score. The column "none" states for each video the number of participants with a 0 score; "full" accordingly the number with the maximum score

Statistic	Mean	St. Dev.	None	Full
Optimal	0.609	0.365	1	2
Random	0.285	0.377	4	0
Single	0.576	0.455	3	3

5.3 Discussion

Several insights are to be gained from the results of our evaluation. First, and most importantly, we can observe a dichotomy between participant preferences and performance. Participants liked the behavior exhibited by the robots in the "random" mechanism video best; however, for that video seem to have remembered the least of what actually happened.

2 Remember that the "like best" ranking is encoded as 1 and "like least" as 3; a more positive ranking for the mentioned questions seems to lead to preferring the video more. Thus, a negative correlation emerges.

Table 4: Summary statistics for the questions regarding people's perception of the robots' explanations; "*o*": "optimal" mechanism video, "*r*": "random" mechanism video, and "*s*": "single" mechanism video

	Min			Max			Mean			St. Dev.			Median		
	o	*r*	*s*	*o*	*r*	*s*	*o*	*r*	*s*	*o*	*r*	*s*	*o*	*r*	*s*
Fake–natural	1	1	2	4	5	4	3.040	3.240	3.000	0.790	1.052	0.707	3	3	3
Machinelike–humanlike	1	1	1	4	4	4	2.320	2.600	2.640	1.030	1.080	1.075	2	2	3
Artificial–lifelike	1	1	1	4	4	4	2.480	2.560	2.760	0.918	1.044	0.879	3	2	3
Static–interactive	1	1	1	5	5	5	2.960	3.200	3.040	1.136	1.258	0.978	3	3	3
Inefficient–efficient	1	1	1	5	5	5	3.360	3.600	3.080	1.075	0.957	1.256	4	4	3
Unpleasant–pleasant	1	2	2	5	5	5	3.560	3.720	3.640	0.917	0.792	0.952	4	4	4
Incompetent–competent	2	3	2	5	5	5	3.800	4.120	3.880	0.707	0.600	0.726	4	4	4
Unintelligent–intelligent	2	2	1	5	5	5	3.360	3.560	3.240	0.907	0.768	1.012	3	4	3

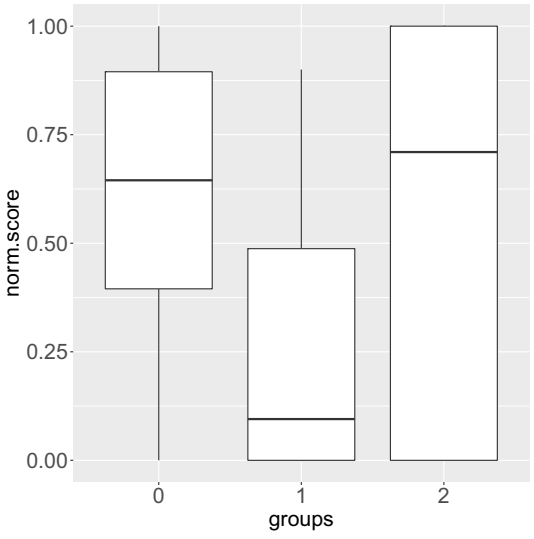

Figure 4: Box plot showing the distribution of memory scores achieved by participants for the different verbalization mechanisms; 0: "optimal," 1: "random," and 2: "single."

While the differences in the memory scores turn out not to be statistically significant – possibly due to a large variety in performance – some differences still seem to be observable. The mean for the "optimal" mechanism group was more than twice as high, and only one participant in this group did not remember anything correctly, compared to four in the "random" mechanism group. Also, two participants achieved a full score, while none managed to do this in the "random" mechanism group, pointing to increased difficulty in remembering events correctly for this group.

In the reproduced plans of the participants in the "optimal" mechanism condition, many of them also included (mostly correct) information on which robot had instructed which other robot to perform an action. None of the participants in the "random" mechanism group did this. We did not account for this in the scoring, since we did not ask for this kind of information in the questionnaire; but it is another clear indication that Grice's maxim of informativeness as followed in constructing these "optimal" explanations increases understanding and memory of what the robots are doing.

The mechanism used in the "single" video seems to lead to performance in between the other two mechanisms but closer to the "optimal" mechanism one. Here, three participants achieved full score but three did not remember anything correctly. The systematic and predictable way of explaining and performing each action separately likely makes it fairly easy to follow what is going on but a little boring to watch; hence, the lowest preference ranking despite good memory performance.

Regarding the rankings along the various behavior scales, it seems that for the "random" mechanism condition, the larger variation in the length of the statements but also the variation in how these cover different, more varied-seeming parts of the plan lead participants to perceive robot behavior to be more interactive and efficient. In fact, two of the participants remarked in their comments that the "random" mechanism video was the only one in which they thought the robots were really interacting. In the others, the robots just seemed to follow "a script" (according to participants' comments). This perception of "script-like" behavior may emerge from the fact that the executed plan is very "linear." With the exception of moving the yellow object aside, the red object is passed on from origin to destination, which is broken into three larger steps executed in strict sequence each by one robot. This may make for very predictable behavior (prediction is to be tested in further empirical studies), which in the "optimal" and "single" mechanism conditions is accompanied by a structured, similarly linear verbalization of what is happening, thereby stating actions only for a single robot at

a time. This might lead to better understandability – in our case better memory for what happened – but also might lead to the robots' behavior being perceived as "machine-like." Since verbalizations in the "random" mechanism are more varied, partly covering actions for several robots, this condition may appear less "script-like" as it may become less predictable. This unpredictability might lead participants to experience the robots to be more competent and intelligent, despite, or maybe rather because of, their behavior being less understandable, as has been shown in previous research (cf. [18]).

Overall, we can conclude that in generating explanations about robot behavior following Grice's maxims, more specifically, optimizing for providing only relevant information, seems to lead to a better understanding in humans about what the robots are doing, but there is also need for further research in order to gain more conclusive answers.

6 Conclusions and future work

In this article, we presented an approach to using policies in distributing tasks among the different robots involved in robot-robot interaction. This approach is flexible in terms of the concrete mechanisms used to perform this distribution and also flexible in that different robots may use different policies in principle. While it is described for a specific scenario, we believe our approach is generally applicable for task distribution and verbalization of actions in many types of sequences of robot actions.

More importantly, we presented an algorithm for partitioning sequences of actions to be used for verbalization in accordance with Grice's maxim of informativeness. This mechanism has been evaluated in a human subject study comparing three different verbalization mechanisms with respect to participants' understanding (memory) of what has happened but also their preferences for these mechanisms. We find some evidence for the benefits of Grice's maxim in terms of understandability; however, there is also a clear preference among the participants for the mechanism picking at random which actions to include in an utterance. This result is in accordance with earlier work [18], showing that unpredictability of a robot increases anthropomorphism and also acceptance of the robot.

The ideas presented in this article have possible extensions along several dimensions. The partitions used for verbalization consist of four information units representing the acting robots, objects to be moved, start, and end point of the objects. The partitions can be generalized to any number of information units representing relevant information for tasks considered. A possible extension is to integrate microplanning and surface realization and to analyze trade-offs regarding the complexity of the given partitions, optimal sequences, policies, and verbalization (i.e., microplanning and surface realization).[3] Furthermore, an incremental approach in which planning and verbalization inform each other and, thus, affect robot understandability is possible. Along the same lines, integrating policy determination in an incremental way may give insight into how the generation of verbal utterances is affected by a chosen policy and how this in the end affects robot understandability.

Finally, there is a need for further empirical evaluation to get a clearer picture of what lies behind the discrepancy between preference and performance as observed in our study. The results of our empirical evaluation point to more work needed to be done to make the explanations generated following Grice's maxims to be perceived as more varied, lifelike, and interactive and, thus, to increase the user acceptance of such explanations.

Acknowledgments: The work presented here is a much-extended version of a previous paper: A. Singh, N. Baranwal, K. F. Richter, T. Hellström, and S. Bensch, "Towards verbal explanations by collaborating robot teams." In International Conference on Social Robotics (ICSR'19), 1st Workshop on Quality of Interaction in Socially Assistive Robots (QISAR'19), Madrid, Spain (2019). This research has been partly funded by the Kempe Foundations and by the Swedish Research Council (Vetenskapsrådet grant 2018-05318).

We thank the anonymous reviewers for their detailed, thorough suggestions to improve this article.

References

[1] T. Hellström and S. Bensch, "Understandable robots – what, why, and how," *Paladyn, Journal of Behavioral Robotics*, vol. 9, pp. 110–123, 2018.

[2] S. Bensch, A. Jevtić and T. Hellström, "On interaction quality in human-robot interaction," *Proceedings of the 9th International Conference on Agents and Artificial Intelligence*, ICAART, Porto, Portugal, vol. 2, pp. 182–189, 2017.

[3] H. Yanco and J. Drury, "Classifying human-robot interaction: an updated taxonomy," in *IEEE International Conference on Systems, Man and Cybernetics*, 2004, pp. 2841–2846.

3 We gratefully picked up a suggestion by one of the reviewers.

[4] M. A. Goodrich and A. C. Schultz, "Human-robot interaction: a survey," *Foundations and Trends in Human-Computer Interaction*, vol. 1, pp. 203–275, 2008.
[5] S. Tellex, T. Kollar, S. Dickerson, M. R. Walter, A. G. Banerjee, S. Teller, and N. Roy, "Approaching the symbol grounding problem with probabilistic graphical models," *AI Magazine*, vol. 32, pp. 64–76, 2011.
[6] L. Baillie, M. E. Foster, C. Breazeal, K. Fischer, P. Denman, and J. R. Cauchard, "The challenges of working on social robots that collaborate with people," in *CHI Conference on Human Factors in Computing Systems* (*CHI'19*), 2019.
[7] A. B. St. Clair and M. J. Matarić, "How robot verbal feedback can improve team performance in human-robot task collaborations," in *HRI '15: Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, 2015, pp. 213–220.
[8] S. Nikolaidis, M. Kwon, J. Forlizzi, and S. Srinivasa, "Planning with verbal communication for human-robot collaboration," *ACM Trans. Hum.-Robot Interact.*, vol. 7, no. 3, pp. 22:1–22:21, 2018.
[9] R. Korpan, S. L. Epstein, A. Aroor, and G. Dekel, "Why: Natural explanations from a robot navigator," in *AAAI 2017 Fall Symposium on Natural Communication for Human-Robot Collaboration*, 2017.
[10] V. Raman, C. Lignos, C. Finucane, K. C. T. Lee, M. Marcus, and H. Kress-Gazit, "Sorry Dave, I'm afraid I can't do that: Explaining unachievable robot tasks using natural language," in *Robotics: Science and Systems*, Technische Universität Berlin, Berlin, Germany, 2013.
[11] A. K. Singh, N. Baranwal, K.-F. Richter, T. Hellström, and S. Bensch, "Understandable collaborating robot teams," in *Highlights in Practical Applications of Agents, Multi-Agent Systems, and Trust-worthiness, The PAAMS Collection*, F. De La Prieta, P. Mathieu, J. A. Rincón Arango, A. El Bolock, E. Del Val, J. Jordán Prunera, et al. Eds., Springer International Publishing, Cham, pp. 168–178, 2020.
[12] H. P. Grice, "Logic and conversation," in *Speech Acts*, ser. Syntax and Semantics, P. Cole and J. L. Morgan, Eds., Academic Press, vol. 3, pp. 43–58, 1975.
[13] K.-F. Richter, *Context-Specific Route Directions – Generation of Cognitively Motivated Wayfinding Instructions*, IOS Press, Amsterdam, the Netherlands, 2008, vol. DisKi 314/SFB/TR 8 Monographs Volume 3.
[14] E. Reiter and R. Dale, *Building Natural Language Generation Systems*, Studies in Natural Language Processing, Cambridge University Press, Cambridge, 2000.
[15] D. H. Younger, "Recognition and parsing of context-free languages in time n3," *Information and Control*, vol. 10, no. 2, pp. 189–208, 1967.
[16] E. Reiter and A. Belz, "An investigation into the validity of some metrics for automatically evaluating natural language generation systems," *Computational Linguistics*, vol. 35, no. 4, pp. 529–558, 2009.
[17] C. Bartneck, E. Croft and D. Kulic, "Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots," *International Journal of Social Robotics*, vol. 1, no. 1, pp. 71–81, 2009.
[18] F. Eyssel, D. Kuchenbrandt and S. Bobinger, "Effects of anticipated human-robot interaction and predictability of robot behavior on perceptions of anthropomorphism," in *HRI '11: Proceedings of the 6th International Conference on Human-Robot Interaction*, 2011, pp. 61–67.