Balancing Shared Autonomy with Human-Robot Communication

Rosario Scalise, Yonatan Bisk, Maxwell Forbes, Daqing Yi, Yejin Choi, and Siddhartha Srinivasa
Paul G. Allen School of Computer Science and Engineering, University of Washington
{rosario, ybisk, mbforbes, dqyi, yejin, siddh}@ cs.washington.edu

Abstract— **Robotic agents that share autonomy with a human should leverage human domain knowledge and account for their preferences when completing a task. This extra knowledge can dramatically improve plan efficiency and user-satisfaction, but these gains are lost if communicating with a robot is taxing and unnatural. In this paper, we show how viewing human-robot language through the lens of shared autonomy explains the efficiency versus cognitive load trade-offs humans make when deciding how cooperative and explicit to make their instructions.**

I. INTRODUCTION

Human-Robot Interaction research often focuses on constructing fully-autonomous systems that work with humans or language-based interfaces that take explicit commands as a form of linguistic teleoperation. Since both the robot and the human have reasoning abilities and unique capabilities, we propose that language can form the basis of a system with shared autonomy wherein both agents help each other accomplish a task by leveraging their individual strengths. In this paper, we investigate how constraints from language balance the cognitive loads both the human and robot need to handle to complete the task, and the language's robustness when scaling the problem size.

Our work views incorporating language with robots through the lens of shared autonomy [1]. Autonomy falls on a spectrum ranging from fully autonomous systems (e.g., manufacturing robots) that require no input from humans, to simple devices (e.g., telescopic cranes) that rely entirely on the user to specify the next action. All scenarios require both reasoning and action. Often we assume both tasks should fall to the robot, but shared autonomy aims to share the burden across all participants. This also allows the user to bias actions of their robotic assistant. This work presents a series of user studies on how humans naturally communicate about a simple table-clearing task to their robot butler, and how knowledge of robotic manipulation capabilities, linguistic comprehension, and task complexity affect a user's language.

Complex environments make planning difficult and automatic plans do not account for user preferences. For example, when packing a suitcase, one might want their toiletries on top for easy access, and they will need to communicate this constraint. When asked for input on plan creation, we find that humans are fast and effective planners, but only once they understand the robot's capabilities. In linguistics, the nature of language a person produces is believed to be guided by Grice's Maxims [2]. We paraphrase two of them here:
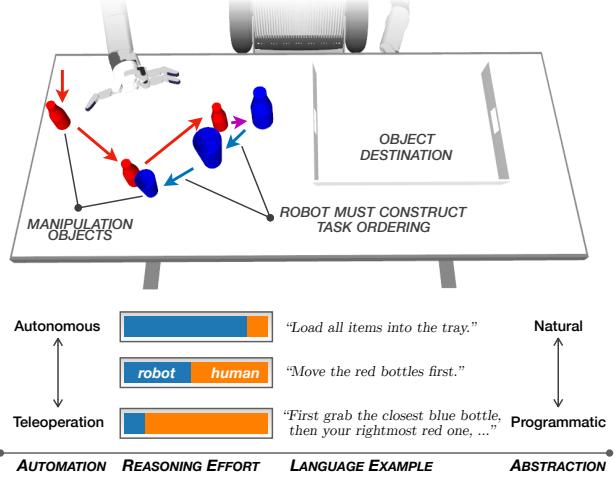

Fig. 1: Shared autonomy is a spectrum along which efforts of both the human and robot trade off in service of a task. We show that a human's expectations of the robot's agency manifest in their utterances.

Quantity: A speaker tries to be as informative as possible while giving only as much information as is necessary.
Manner: A speaker tries to be brief and to avoid ambiguity.

What is most important to understand about these principles is that they are fundamentally social. To know what or how much to say requires modeling the listener. In the case of robotics, most users have either a weak or incorrect model of the robot's abilities and so, without guidance, they make the wrong assumptions about what is necessary and what is useful to communicate. We explore this later in the paper.

We hypothesize that language varies from natural to programmatic (Fig. 1) depending on a user's expectations of the robots capabilities. Our second hypothesis is that if asked to, humans are effective and efficient planners whose insights, preferences and task intuitions can be effectively harnessed to improve task performance. Our experiments use the task and motion planning library, MAGI [3], to translate the high-level table-clearing task into an ordering of table-top objects and the corresponding motion plans needed to execute moving each object.[1] We setup our experiments within this planning framework to allow users to specify high-level goals or orders of objects to manipulate. We avoid bias that might be introduced by working with members of the university community by running studies on Amazon's Mechanical Turk

---

[1]The planner performs a depth-first search over a permutation graph of manipulable objects until it finds a geometrically feasible solution which addresses all objects.

(AMT) with participants who had no previous knowledge of robotics. We constrain the experiments to a simple table clearing task [3] but feel these results generalize to other domains and robotic platforms.

We first describe a preliminary study (§III) which guides the creation of our AMT study on comparing plan efficiency and user effort (§V). Next, we investigate a new type of language to the literature that occurs in complex domains (§VI) and analyze the language strategies users produced (§VII).
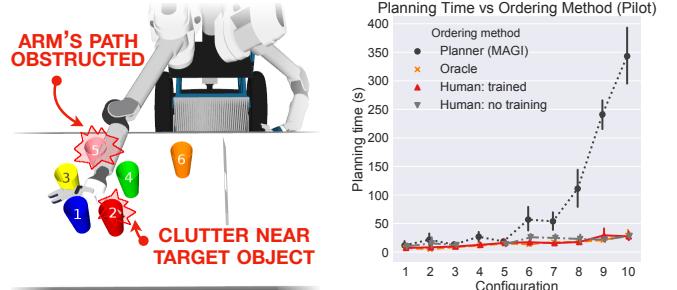
II. RELATED WORK

Our work is motivated by the rich breadth of research on using natural language to communicate with robots both within the robotics and natural language communities. We are interested in how language interaction shifts the cognitive load in shared autonomy.

Previous work has grounded natural language navigational commands to executable representations. Graphical model-based approaches using syntactic parses have been applied to controlling robotic forklift actions [4] and mobile navigation in novel environments [5], [6]. Others have utilized CCG semantic parsing of robotic commands in synthetic environments [7], and with weak supervision [8]. Howard (2014) [9] parsed natural language to constraints in trajectory space in order to reduce the search space of their graphical model that grounds language to instructions. Park (2017) [10] used a similar approach, grounding instructional language to a probabilistic graphical model, though they address the manipulation domain and learn soft cost functions to avoid dynamically-specified regions (e.g., "don't put it there").

Language interaction provides useful information in solving many of the challenges in shared autonomy, which include how to correctly and accurately identify a human's intent through interaction [1] and observation [11]. Low-level control strategies from humans have also been successfully integrated and applied to search terrain [12]. Recent research aggressively incorporates high-level human information in shared autonomy [13], [14]. Language, which is a direct and natural way for a human to share information in collaboration, has been widely researched within the "supervisor" paradigm [15], [16], [9] where only a goal is provided or where the human acts as a "programmer" [17], [18], [19] that instructs the agent in tedious detail. Balancing these roles and investigating user preferences is still an open challenge for deploying communication in shared autonomy.

Finally, when referencing ambiguous objects, people use visual attributes and referring expressions to indicate which object should move next. Like our work, approaches to referring expressions largely focus on the tabletop domain [16], and offer systems that are interactive [15], collaborative [20], or incorporate manipulation actions [19]. Referring expressions can also be used as a means of effectively asking for help [21]. Finally, the Natural Language Processing (NLP) community has recently also introduced corpora for this task with substantially more complex linguistic constructions [22]. Given such varied language between papers, our

(a) Low-level motion planning might not find a collision-free solution to a multi-step manipulation task if forced to follow a bad action order.

(b) Planning time increases significantly when the task is complex, but incorporating human insight can offer dramatic performance gains.

work aims to better understand what leads a user to choose a given communicative strategy.

III. PRELIMINARY STUDY

Our preliminary study investigates whether humans are good planners and how their language changes when they are taught about a robot's capabilities and planner. The training we provide partway through the study is a proxy for the online learning they would receive when interacting with a robot, but allows us to collect multiple data points for an individual user. This provides us a clear before-and-after comparison to investigate the effect of training.

We explained the goal task (table-clearing) to the user and that our robot, HERB, will only be using his right arm to manipulate one object at a time (partially replicated below). We then provided ten object configurations images like those shown in the top box of Figure 3 (the design rational of which is explained in greater detail in section IV) to nine participants and requested instructions one-by-one:

> *You are overseeing a robot (HERB)s role as a butler, ensuring he completes his tasks correctly and efficiently. Occasionally, he needs a little guidance.*
> *...*
> *For each scenario below, write an instruction for HERB to follow. Remember, all cups must end up in the tray.*

A. Untrained

For the initial experiment no additional information was provided. Unsurprisingly, most participants provided vague goal-oriented language:

1) *Place all cups in the tray*   `[goals]`
2) *Pick up the closest cup and move it onto the tray. Repeat until there are no more cups.* `[algorithmic]`

Even a user who anticipated the fine-grained planning needs of the robot and initially provided verbose step-by-step instructions like: *"Pick up the red cup and put it in the tray, then pick up the blue cup and put it in the tray, then..."* eventually switched to vague phrasings after growing bored with the task.

As it is clear that these "plans" contain no immediately useful information for the planner, once all scenarios were completed we asked participants to "translate" their own

instructions into an explicit ordering over subtasks corresponding to each object on the table. Here, they are making concrete their own assumptions about how to complete the task. We pass these orderings to the planner and compare its planning times to those produced by vanilla planner runs (meaning it has no priors on ordering over subtasks for objects on the table). Plan performances are shown in Figure 2b. The ten instructions cover tables with three through seven uniquely colored cups for both densely and loosely cluttered scenes. We chose these to compare the breadth of difficulty for the planner as compared to the human participants.

B. Training

Next, we train users on basic details of planning algorithms in colloquial terms and with a technical definition:

> HERB must sample and evaluate trajectories to assess if he can reach the cup and manipulate it. The fewer viable paths, the more samples are rejected and the harder it is to plan successfully.

Alongside this explanation we provide the user a basic demonstration of what to avoid when giving an ordering (Figure 2a). We do not instruct users in how to communicate. After training, we repeat our data collection with ten new prompts. This brief training leads participants to converge on the use of long verbose plans with explicit orderings.

C. Preliminary Experimental Results

This preliminary study shows that a small amount of training dramatically affects robot-oriented language in a way that's easier for the robot but less natural and more cumbersome for the person (quantitative metrics and analysis of our large-scale studies are presented in section V). Next we evaluate the plans users produced to see how well human intuitions map to good motion plans.

Figure 2b suggests two important findings. First, humans intuit good orderings well (outperforming our baseline in planning time), and second, even untrained users perform well. This might indicate that their initial high-level instructions assumed a very capable planner, and when training introduced doubts they made their assumptions explicit.

IV. STUDY DESIGN

A. Stimulus Design

We programatically generated stimulus images which consisted of colored cups (and/or) bottles in varying configurations on a table top. Each of the images was rendered from a 45 degree inclination angle and had the robot, HERB, present sitting in a pre-task configuration behind the table from the viewers perspective. Each of the configurations could be varied by the number of objects in the scene, whether the objects were sparsely placed or packed tightly, and by the object attributes available (object type, color, and size). When scaling the number of objects in a configuration, we chose to incrementally add an additional object to an existing set to create the new configuration. Doing this allowed us to be confident that there were no additional factors that could affect the changing complexity

Fig. 3: Example configurations and language from both experiments. The orange boxes denote the type of language used in the shown example.

between two configurations.

Experiment I Stimuli Our first experiment consists of two types of stimuli (Figure 3). The first set of stimulus images was composed of 5 'sparse' configurations and 5 'packed' configurations (each of which contained between 3 and 7 uniquely colored large-sized cups) for a total of 10 images.

The second set of images was composed of 6 'sparse' configurations and 6 'packed' configurations where the number of objects for all configurations was fixed at 6, colors were restricted to only red and blue (no longer unique), and objects' attributes were systematically varied. We generated a 'sparse' and a 'packed' configuration for each possible combination of available discriminating object attributes. For example, in the first two configurations, color is the only available attribute while in the last two configurations, color, type, and size are all varied among the objects. Our rational for picking 6 objects and not more was that this resulted in sufficient complexity without creating longer run-times for our planner. In total, the second set contained 12 images.

Experiment II Stimuli Our second experiment investigates the effect of scaling and complexity. Here the set of stimulus images was composed of 6 configurations with 24 uniquely colored cups and 6 configurations of 24 objects in which they were randomly assigned attributes (a color from $red, blue$, a size from $small, large$, and a type from $cup, bottle$). In total, the third set contained 12 images (examples in Fig. 3).

B. Subject Allocation

Both Experiments I and II were deployed via AMT. We recruited a total of n=50 participants for each, ensuring that participants who had seen Experiment I were not eligible to do Experiment II. We required that each participant was a native English speaker and was not color blind. We also surveyed participants on their past experience with robots at the end of each study.

V. EXPERIMENT I

The results of the preliminary study corroborate our intuition that researching human robot collaboration in decision making falls within the framework and goals of shared autonomy. Specifically, we noted that humans are good at high-level reasoning and can specify an efficient ordering over subtasks in a high-level planning task. However, this shifts the cognitive workload to the human which they are reluctant to accommodate. In contrast, robots are unlikely to have context specific heuristics about the environment and therefore have more difficulty finding a good orderings over subtasks, but they are very good at low-level path-planning for generating motion trajectories for individual subtasks and are highly capable assistants. This asymmetry between humans and robots can be generalized to a wide range of shared autonomy systems. Both agents will always have different capabilities or specialties, and shifting the cognitive load required by the human biases the system in one direction on the spectrum of full-autonomy to full teleoperation.

Understanding the breadth of language people tend to use in these interactions helps to

1) design language understanding algorithms and language communication schemes to support human-robot communication;
2) enable a robot to interpret the current intent of a human in sharing the workload; and
3) allow a robot to actively generate language in order to shift its contribution on the autonomy spectrum to optimize the team performance.

Our first experiment is designed to test two hypotheses:

H1 *Humans tend to provide "natural" expressions when they trust in the robot's capabilities (they assume high robot autonomy before any information is given), which requires less work (mental + temporal demands).*

H2 *Humans tend to provide "programmatic" expressions when they are aware of a robot's limitations, despite it increasing their workload.*

A. Study Design

We use the stimuli discussed in §IV in our experiment on AMT. Training is split into two phases: 1. **R**: Robot capabilities and 2. **R+C**: Information about language phrasings that the robot can understand (constraints and orderings). This two phase approach ensures that our "translation" step from the preliminary study cannot bias participants towards listing objects.

B. Evaluation

To test our hypotheses we perform three evaluations based on plan efficiency, participant self-assessment, and linguistic analysis.

Plan Efficiency We annotate all plans provided by users into orderings (or partial-orderings) over the manipulable objects in the scene. With these orderings, we run the motion planner and assess each plan's efficiency by averaging over 10 plan times. When users provide high-level instructions (which contain no ordering information), we use the planner's de facto performance on the task. When users provide instructions which map to partial-orderings, we average the planning time of 3 samples from the set of feasible orderings which satisfy the given partial ordering.

Participant Self-Assessment We ask our participants two categories of question: 1. We ask about the effort involved in the task and 2. We ask if they prefer scenes with simple referents or sets of objects. User effort is collected with questions based on NASA TLX [23]. For each instruction we ask:

1) *How mentally demanding was it to come up with this?*
2) *How much time did it take you to think of this?*
3) *How satisfied are you with your performance on this?*

Linguistic Analysis We introduce two metrics for how the language of our participants vary through several conditions. First, we analyze type-token ratios. Natural language is highly varied and rarely formulaic. This manifests in a high type-token ratios (number of unique word types vs number of total words used). In contrast, when language becomes repetitive and programmatic the ratio will fall. We therefore present type-token ratios for all conditions as a proxy for how programmatic and "unnatural" language has become.

Second, we manually annotated all of the responses as falling into one of three categories: *high-level*, *partial ordering*, and *absolute ordering*. We therefore report the number of instructions in each condition which belong to these classes.

C. Results

Our primary interest is in understanding the trade-off between the mental energy exerted by the participant versus their informativeness to the planner.

1) Language Analysis: H1 and H2 supposed that language would change in potentially drastic ways when users are trained on a robot's capabilities. The left column of Figure 4 shows precisely this effect. On top, we show how users were split between different linguistic approaches, but once trained they shift almost exclusively to absolute orderings. It is only after we tell them that constraints and partial orderings are acceptable that they begin to change their approach.

In Figure 4 (below), we see a very similar trend in the type token ratios. Again, untrained users take a diverse set of strategies but once trained they shift towards very repetitive language (low type/token ratios). Given these jarring changes
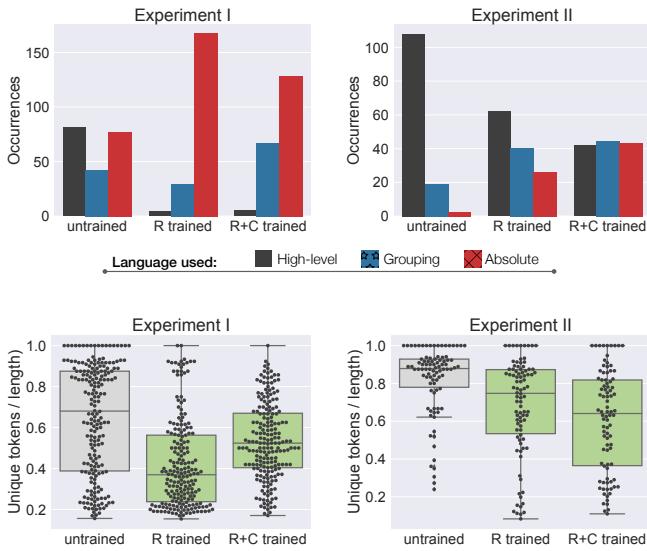
Fig. 4: Language analysis. **Top:** After training users in robot capabilities, we observe an increase in absolute ordering language. In Experiment I, we also observe an increase in grouping language after communication training (§V-A). **Bottom:** Word repetition analysis (1.0 = no repetition, 0.0 = complete repetition). We observe that after robot capability training, repetitiveness increases, though this is then offset in Experiment I with communication training. In Experiment II, users overall relied more heavily on algorithmic language due to the complexity of the task.
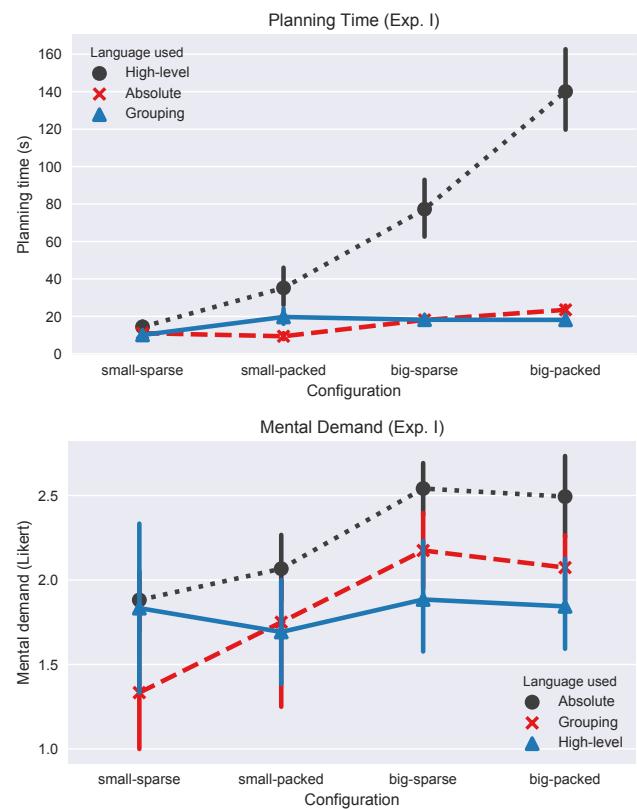

Fig. 5: **Top:** High-level language does not assist the planner, but both grouping language and absolute orderings provide considerable gains. **Bottom:** Grouping language strikes a balance in mental demand. Overall, grouping language limits mental demand while benefiting planner performance.

to the language, we assess their effects on plan efficiencies and correlation with mental demand.

2) Plan Efficiency vs Mental Demand: Every type of language (shown in Fig. 3) provides an incrementally larger amount of information to the planner. For every configuration, we ran our planner 10 times on the most common constraint/ordering provided (error bars shown in figure). Figure 5 shows the result of these instructions in planning time. For analysis, we aggregate configurations into four categories: Packed vs Sparse, and Small (≤ 5) vs Large (≥ 6). When only a high-level plan is provided, we default to the planner's de facto performance. We use exact sequences for absolute orderings. Finally, if only a partial ordering is provided, we average over orderings sampled from the set of orderings that lead to geometrically feasible solutions. Immediately, we see that human insight dramatically speeds up planning.

Next, we plot the mental demand users indicated was required for each type of language and configuration. First we note that high-level plans are unaffected by task complexity while providing absolute orderings becomes tiring. Importantly though, our results for grouping language which only partially constrain the task validate H2, as they balance plan efficiency (providing large gains to the planner) while incurring a lower mental demand than fully explicit orderings.

3) Mental Exhaustion: Finally, we investigate how the phases of our task are wearing on the participants. The top of Figure 5 shows how, regardless of the scenario, users are very satisfied with their performance but the time and mental energy they are spending on the task increases and decreases as they become aware of the robot's limitations and linguistic abilities, respectively. Their desire to communicate effectively but easily also shows up as a strong preference for scenes in which all cups have a unique color (Figure 7). This makes sense since naming a color is easier than using spatial language or multiple attributes to disambiguate an object ("red cup" vs "red cup on the left").

VI. EXPERIMENT II

Experiment I limited scenarios to at most seven objects to be consistent with most robots+language literature which stay under 15 objects. In contrast, most natural scenes are more ambiguous with a large number of possible referents. While complex environments are exponentially more difficult for motion planners, in our simulated environment, we can scale our experiment to see the effects on language. This leads us to a new hypothesis untested in the literature:

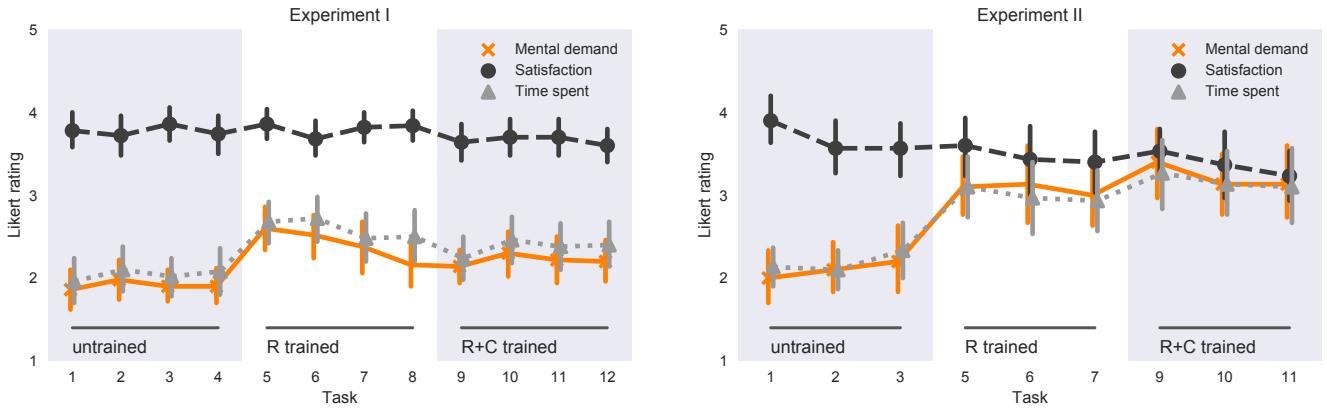**H3** *The preference for programmatic language and*

Fig. 6: Mental demand, satisfaction, and time ratings for both experiments. Difficulty increases after robot training and language training, but appears to level off as users acclimate to the task.

absolute orderings is an artifact of people's desire to minimize their own cognitive load and therefore an artifact of simple environments. Specifically, we expect constraint and set based language to be most common in complex environments.

A. Study Design

Our design mirrors Experiment I but uses a new set of stimuli which focuses on scaling the number of objects in a scene. These were divided into two scene types with 24 objects in specified locations. In the first, each of the 24 objects was uniquely identifiable by a color with a common name. In the second, each of the 24 objects was randomly assigned a value for each attribute (bottle/cup, blue/red, small/large). Again, we disallow any user overlap from the previous experiment, so all participants are new to the task and untrained.

B. Results

While the planner will now be too slow to compare as a possible baseline, we still want to investigate how language and mental exhaustion scale to more realistic scenarios.

1) Language Analysis: Where simple environments occasionally elicited explicit ordering language even before training, these are not seen as natural or viable approaches in the richer environment. Figure 4 (top) shows a dramatic preference for high-level language, and correspondingly, we see very high type/token ratios (bottom). These are truly natural and diverse instructions. More importantly, and interestingly, we see a remarkable reluctance to change even after training. Where knowledge of robot capabilities previously led participants to use the very helpful (though taxing) absolute orderings, now users choose to only provide partial orderings or set-based language. It is only after we explicitly tell them of the types of language we want that absolute orderings become more common. Even then, they are half as common as in the simpler setting. Again, the lowest type-token ratios in this experiment are higher on average than Experiment I.
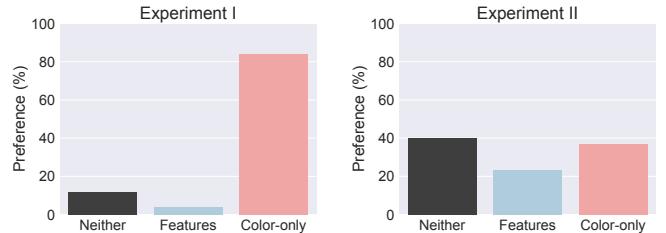

Fig. 7: Preference ratings across task configurations for Experiments I and II. In Experiment I, users preferred using simple referring expressions to locate unique objects. Users did not demonstrate any preference in Experiment II.

2) Mental Exhaustion: Another important differentiation of this result from Experiment I is seen in the changing mental demands. Figure 5 shows how, after training, mental demand increases and never drops. Where before, we gave participants a "way out" by suggesting they try and use constraints or orderings to simplify the communication, now they are reluctant to change their language and forcing them to consider plans in detail is highly demanding. Particularly poignant is that while demand stays constant, satisfaction decreases. Users are increasingly uneasy about their own instructions and how they will be interpreted. This leaves open some important questions about how to decide when it is worth pressing the human for their powerful insights if they find the process frustrating.

Correspondingly, now that the scenes are complicated, preferences for easily referenced colors versus sets even out. Figure 7 shows that unique colors are still preferred, but there is little agreement as to which setting is actually easiest.

VII. Language Strategies

Upon publication we will make available all of our configurations and the corresponding language for the community to scrutinize, but we include a few examples here in Figure 3. We want to draw particular attention to the different types of high-level language. Recent results in robotics lead us to believe that many labs can handle absolute orderings and

recent work on understanding groups/rows/sets should cover partial orderings, but high-level language appears to be much less homogeneous in nature. Specifically, in Experiment II the majority of users tried to be helpful by providing us with techniques, strategies, and intuition for the problem.

These example differ from goal-language as they are closer to pseudocode for the correct search procedure. In our experiments, HERB's abilities nicely parallel a human's arm and so the user might be describing how they would reason about the task. Equally important for future work is to include details subtle ways the robot differs from their expectations (e.g. HERB's long arms might make close grasps difficult), and then compare the algorithms/heuristics generated by the user. More technically, we are unaware of any literature that works to interpret and convert these types of hints into planner actions or constraints.

Finally, our participants all tried to be helpful to HERB. When we compared the most common plan (the ones used for computing motion plans in our plots) to all others produced within the pilot, we saw on very small differences in plan time since nobody strategically instructed HERB to perform infeasible actions. This may not be true in general for deployed robotics, hinting at a new research question: How do we detect when a user is being malicious?

VIII. Conclusion

This work discusses the interaction between humans and robots from a language communication perspective. It investigates the importance of language in shifting autonomy between the human and robot, and when or why a human might choose to be helpful or abdicate responsibility.

We only discussed the language from human to robot in this work, which allows the human to decide how much autonomy they want the robot to exhibit. Understanding the variables we introduced helps us calibrate the shared load a human teammate expects. Analysis and categorization of language also provides insight into how much of the workload the human teammate is willing to take-on or how hard they are working on a given task. A natural extension is to inquire how a robot should respond (verbally) if they want to strategically ask for help or increase the user's participation to redistribute or optimize the cognitive load.

Humans work very well with each other in teams by communicating goals, plans, heuristics, and asking for help. While semantic parsers and Natural Language Processing (NLP) systems may not yet be equipped to handle all of the parsing necessary for the language we have presented, if robots are to serve as helpful team members, we will need to bridge this gap between human language preferences and robot understanding ability.

This is still a young area of research, and language communication in shared autonomy provides an exciting and effective interface to enhance existing sliding autonomy systems [24] into complex task coordination, where-in task-planning libraries can be extended to be interactive task-planning libraries that elicit language interaction. We believe these research results will help us to better design agents with

bidirectional communication between humans and robots, especially in manipulation tasks, where a robot needs to: (1) precisely understand what a human wants, (2) dynamically monitor the workload distribution, and (3) model the human's characteristic behaviors for optimizing reactions.

References

[1] S. Javdani, S. S. Srinivasa, and J. A. Bagnell, "Shared autonomy via hindsight optimization," *arXiv preprint arXiv:1503.07619*, 2015.

[2] R. Frederking, "Grices maxims: do the right thing," *Frederking, RE*, 1996.

[3] S. Srinivasa, G. Johnson, A.and Lee, M. Koval, S. Choudhury, J. King, C. Dellin, M. Harding, D. Butterworth, P. Velagapudi, and A. Thackston, "A system for multi-step mobile manipulation: Architecture, algorithms, and experiments," in *International Symposium on Experimental Robotics*, 2016.

[4] S. Tellex, T. Kollar, S. Dickerson, M. R. Walter, A. G. Banerjee, S. J. Teller, and N. Roy, "Understanding natural language commands for robotic navigation and mobile manipulation." in *AAAI*, 2011.

[5] S. Hemachandra, F. Duvallet, T. M. Howard, N. Roy, A. Stentz, and M. R. Walter, "Learning models for following natural language directions in unknown environments," in *Robotics and Automation (ICRA), 2015 IEEE International Conference on*. IEEE, 2015, pp. 5608–5615.

[6] D. Yi, T. M. Howard, M. A. Goodrich, and K. D. Seppi, "Expressing homotopic requirements for mobile robot navigation through natural language instructions," in *Intelligent Robots and Systems (IROS), 2016 IEEE/RSJ International Conference on*. IEEE, 2016, pp. 1462–1468.

[7] C. Matuszek, E. Herbst, L. Zettlemoyer, and D. Fox, "Learning to parse natural language commands to a robot control system," in *Experimental Robotics*. Springer, 2013, pp. 403–415.

[8] Y. Artzi and L. Zettlemoyer, "Weakly supervised learning of semantic parsers for mapping instructions to actions," *Transactions of the Association for Computational Linguistics*, vol. 1, pp. 49–62, 2013.

[9] T. M. Howard, S. Tellex, and N. Roy, "A natural language planner interface for mobile manipulators," in *Robotics and Automation (ICRA), 2014 IEEE International Conference on*. IEEE, 2014, pp. 6652–6659.

[10] J. S. Park, B. Jia, M. Bansal, and D. Manocha, "Generating real-time motion plans from complex natural language commands using dynamic grounding graphs," *arXiv preprint arXiv:1707.02387*, 2017.

[11] A. D. Dragan, K. C. Lee, and S. S. Srinivasa, "Legibility and predictability of robot motion," in *Human-Robot Interaction (HRI), 2013 8th ACM/IEEE International Conference on*. IEEE, 2013, pp. 301–308.

[12] Y. Okada, K. Nagatani, K. Yoshida, S. Tadokoro, T. Yoshida, and E. Koyanagi, "Shared autonomy system for tracked vehicles on rough terrain based on continuous three-dimensional terrain scanning," *Journal of Field Robotics*, vol. 28, no. 6, pp. 875–893, 2011.

[13] N. R. Ahmed, E. M. Sample, and M. Campbell, "Bayesian multicategorical soft data fusion for human–robot collaboration," *IEEE Transactions on Robotics*, vol. 29, no. 1, pp. 189–206, 2013.

[14] S. C. Daqing Yi and S. Srinivasa, "Incorporating qualitative information into quantitative estimation via sequentially constrained hamiltonian monte carlo sampling," in *Intelligent Robots and Systems (IROS), 2017 IEEE/RSJ International Conference on*. IEEE, 2017.

[15] T. Kollar, J. Krishnamurthy, and G. P. Strimel, "Toward interactive grounded language acqusition." in *Robotics: Science and systems*, 2013.

[16] C. Matuszek, N. FitzGerald, L. Zettlemoyer, L. Bo, and D. Fox, "A joint model of language and perception for grounded attribute learning," *arXiv preprint arXiv:1206.6423*, 2012.

[17] T. Sato and S. Hirai, "Language-aided robotic teleoperation system (larts) for advanced teleoperation," *IEEE Journal on Robotics and Automation*, vol. 3, no. 5, pp. 476–481, 1987.

[18] S. Lauria, G. Bugmann, T. Kyriacou, and E. Klein, "Mobile robot programming using natural language," *Robotics and Autonomous Systems*, vol. 38, no. 3-4, pp. 171–181, 2002.

[19] M. Forbes, R. P. Rao, L. Zettlemoyer, and M. Cakmak, "Robot programming by demonstration with situated spatial language understanding," in *Robotics and Automation (ICRA), 2015 IEEE International Conference on*. IEEE, 2015, pp. 2014–2020.

[20] R. Fang, M. Doering, and J. Y. Chai, "Embodied collaborative referring expression generation in situated human-robot interaction," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 2015, pp. 271–278.

[21] S. Tellex, R. A. Knepper, A. Li, D. Rus, and N. Roy, "Asking for help using inverse semantics." in *Robotics: Science and systems*, vol. 2, no. 3, 2014.

[22] Y. Bisk, D. Yuret, and D. Marcu, "Natural language communication with robots." in *HLT-NAACL*, 2016, pp. 751–761.

[23] S. G. Hart, "Nasa-task load index (nasa-tlx); 20 years later," in *Proceedings of the human factors and ergonomics society annual meeting*, vol. 50, no. 9. Sage Publications Sage CA: Los Angeles, CA, 2006, pp. 904–908.

[24] F. W. Heger and S. Singh, "Sliding autonomy for complex coordinated multi-robot tasks: Analysis & experiments," 2006.