COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 47 (2023) e13396

ISSN: 1551-6709 online
DOI: 10.1111/cogs.13396

The Efficiency of Question-Asking Strategies in a Real-World Visual Search Task

Alberto Testoni,[a] Raffaella Bernardi,[b,c] Azzurra Ruggeri[d,e,f]

[a]*Institute for Logic, Language and Computation (ILLC), University of Amsterdam*
[b]*Center for Mind/Brain Sciences (CIMeC), University of Trento*
[c]*Department of Information Engineering and Computer Science (DISI), University of Trento*
[d]*MPRG iSearch, Max Planck Institute for Human Development, Berlin*
[e]*School of Social Sciences and Technology, Technical University Munich*
[f]*Department of Cognitive Science, Central European University*

Received 15 December 2022; received in revised form 14 November 2023; accepted 1 December 2023

Abstract

In recent years, a multitude of datasets of human–human conversations has been released for the main purpose of training conversational agents based on data-hungry artificial neural networks. In this paper, we argue that datasets of this sort represent a useful and underexplored source to validate, complement, and enhance cognitive studies on human behavior and language use. We present a method that leverages the recent development of powerful computational models to obtain the fine-grained annotation required to apply metrics and techniques from Cognitive Science to large datasets. Previous work in Cognitive Science has investigated the question-asking strategies of human participants by employing different variants of the so-called 20-question-game setting and proposing several evaluation methods. In our work, we focus on GuessWhat, a task proposed within the Computer Vision and Natural Language Processing communities that is similar in structure to the 20-question-game setting. Crucially, the GuessWhat dataset contains tens of thousands of dialogues based on real-world images, making it a suitable setting to investigate the question-asking strategies of human players on a large scale and in a natural setting. Our results demonstrate the effectiveness of computational tools to automatically code how the hypothesis space changes throughout the dialogue in complex visual scenes. On the one hand, we confirm findings from previous work on smaller and more controlled settings. On the other hand, our analyses allow us to highlight the presence of "uninformative" questions (in terms of Expected Information Gain) at specific rounds of the dialogue. We hypothesize that these questions

Correspondence should be sent to Alberto Testoni, Institute for Logic, Language and Computation (ILLC), University of Amsterdam, Science Park 107, 1098 XG Amsterdam, Netherlands. E-mail: a.testoni@uva.nl

fulfill pragmatic constraints that are exploited by human players to solve visual tasks in complex scenes successfully. Our work illustrates a method that brings together efforts and findings from different disciplines to gain a better understanding of human question-asking strategies on large-scale datasets, while at the same time posing new questions about the development of conversational systems.

Keywords: Visual search; Information search; 20-Questions game; Question asking; Expected information gain

1. Introduction

The last few years have witnessed significant progress in developing conversational agents capable of successfully interacting with humans in everyday-life situations (Singh & Beniwal, 2022). These agents are based on complex data-hungry artificial neural network architectures that require large amounts of human–human conversations to process and generate human-like language, at least on a surface level. To this aim, several datasets of human participants carrying out a wide variety of different tasks have been collected and released. In order to successfully train an artificial neural network model, an impressive amount of conversations is required, ranging from tens of thousands to millions of examples. Crowdsourcing has become the standard practice to collect linguistic data of this sort, where anonymous workers perform tasks in exchange for a small amount of money. To make the dataset collection easier and elicit natural conversations, the task under analysis is usually framed within a gamification setting between two participants. In this paper, we argue that datasets of this sort represent a unique resource for cognitive scientists to study human behavior and language use on a large scale, while at the same time providing insights into the critical features of human language that are essential when developing effective and proficient conversational agents. We believe that question-asking and visual search tasks are particularly suited for studies of this sort, given the recent availability of large datasets and the long-standing body of work of cognitive studies investigating these tasks in smaller and more controlled settings. The research community working on Natural Language Processing (NLP) has proposed many complex tasks that require computational models to learn how to ask informative and strategic questions to reach a goal. The main focus of this research direction is on the engineering side of developing complex artificial neural network architectures for the task at hand, while the analysis of such data is often left unexplored. We believe a Cognitive Science perspective could provide useful and much-needed insights to support the development of more sophisticated conversational agents. As most human language understanding is grounded in perception, this has recently become particularly relevant to NLP and Computer Vision researchers who are working on developing language-driven visual models that require a plethora of different skills.

1.1. Conversational agents and linguistic competence

On the engineering side, much effort has been devoted to developing sophisticated and accurate encoder models for Natural Language Understanding tasks, that is, models that take as input human-generated text and try to understand the user's requests. Recently, encoders

have been paired with decoders that learn to generate text resembling human language (Sordoni et al., 2015). As a result, virtual assistants can (most of the time) accurately process users' inquiries. However, their competence is often limited to tasks where the virtual assistant is conceived as a passive agent that simply has to satisfy the user's requests. Voice assistants are so far capable to ask questions to clarify the user's requests (e.g., "Do you mean…?"), but they do not display what can be considered intelligent problem-solving behavior. Crucially, they output fluent natural language but they lack reasoning skills behind the generation of utterances. As argued in Mahowald et al. (2023), even Large Language Models such as ChatGPT[1] that successfully master "formal linguistic competence," defined as knowledge of rules and patterns of a given language, fall short on "functional linguistic competence," that is, the set of non-language-specific cognitive abilities required for modeling reasoning and thought in the real world. Cooperation between computer scientists and cognitive scientists is essential to shed light on the unique features of human language use that are required for developing conversational agents that can assist humans more effectively in daily real-world scenarios.

1.2. Asking informative questions as a cross-disciplinary problem

The ability of computational models to ask a sequence of informative questions to reach a goal is clearly of interest to different research fields, which so far have not experienced much cross-pollination. On the one hand, a long tradition in Cognitive Science has investigated how humans learn to actively solve problems and ask questions to gather information about the state of the world and to achieve a goal. This line of work has often employed variants of the *20-question* game, in which participants are tasked to identify an unknown target object by asking as few yes-or-no questions as possible (Meder & Nelson, 2012; Nelson, Divjak, Gudmundsdottir, Martignon, & Meder, 2014; Ruggeri, Lombrozo, Griffiths, & Xu, 2016; Ruggeri & Feufel, 2015). In these tasks, a key role is played by the *informativeness* of the questions asked, which can be computed in different ways (e.g., success, probability gain, impact, expected savings, and expected information gain [EIG]; for a review, see Nelson, 2005), and can be modeled within a Bayesian framework originally developed for concept learning and generalization (Tenenbaum & Griffiths, 2001). Work using this paradigm has identified three milestones in the developmental trajectory of children's question-asking strategies: at 5, 7, and 10 years of age. Children's question-asking abilities evolve from being able to identify good questions but not being able to spontaneously generate them at the age of 5 (e.g., Ruggeri, Sim, & Xu, 2017; Ruggeri, Walker, Lombrozo, & Gopnik, 2021), to beginning to generate them spontaneously at age 7 (see Ruggeri & Feufel, 2015; Herwig, 1982), and implementing efficient and adaptive question-asking strategies by the age of 10, echoing adult-level patterns of performance. Moreover, this line of work has shown that participants' performance outperforms random search, but systematically falls behind optimal performance (Meder & Nelson, 2012; Ruggeri et al., 2016; Rothe, Lake, & Gureckis, 2018). This observation combines with the findings from Gatt, van Gompel, van Deemter, and Krahmer (2013), who showed that when referring to objects with different features in a visual scene, a "rational" Bayesian model, whose choices are based on utility, cannot account for

participants' preferences: Humans tend to be overspecific and prefer properties irrespectively of their utility for identifying the referent (e.g., referring to the color of the target even when this is not a discriminative feature against the other objects). Their choices are instead guided by simpler heuristics, relying on perceptual or cognitive salience. In particular, in line with previous work (Belke & Meyer, 2002; Eikmeyer & Ahlsén, 1996), the authors considered a set of referents of the same category which differ from each other in size and color: In trials where the size alone was sufficient to uniquely identify the target among the other objects, human speakers show a strong preference for mentioning the color of the target, even if this property is shared across multiple referents.

1.3. Visual search

Many of the works discussed above on investigating human question-asking strategies are framed as referential tasks where participants have to locate a target object in visual scenes. Identifying a target in a visual scene is a core ingredient of many daily tasks (Chan & Hayward, 2013; Eckstein, 2011), from searching for the remote control in a messy room, to looking for a familiar face in a crowded street, to finding something to eat in the kitchen. In this sense, visual search strategies are a fundamental component of human intelligence: They can be guided by a single individual with a goal in mind, or they may result from interaction with an interlocutor. For instance, visual search guided by the cooperation with another person is crucial when there is information asymmetry between two partners (e.g., the target is known only to the other person), or when assisting and guiding another person in a novel or challenging task (e.g., navigating a new environment). Given its practical and theoretical implications, over the last couple of decades, modeling visual search has attracted the attention of research communities and industries working on developing the next generation of virtual assistants. Visual search skills are particularly relevant for conversational agents designed to help visually impaired and elderly people with daily tasks. Imagine a person who has difficulty moving and asks a robot to grab objects around them, or to get information about a particular item (e.g., the expiring date of some food or medicine, see, for instance, the dataset released in Gurari et al., 2018). Developing agents that can ask a sequence of effective questions about the surrounding environments to locate objects and resolve ambiguities about the users' inquiries are crucial steps in this direction. More generally, there is very limited behavioral work analyzing humans' information search behavior in visual tasks. Previous work has investigated the role of visual grouping in visual search tasks, exploring the effect of spatial configurations on the search for objects and features in scenes containing clusters of objects (Treisman, 1982). Fixation eye movements can offer some insights into understanding the mechanisms underlying visual search. For example, some work has explored the role of top-down (i.e., guided by the task/goal of the observer) (Hayhoe & Ballard, 2005; Hayhoe, Shrivastava, Mruczek, & Pelz, 2003; Land, Mennie, & Rusted, 1999; Yarbus, 1967) and bottom-up (i.e., stimulus-driven) factors in fixational eye movements (Kundel & La Follette Jr, 1972; Parkhurst, Law, & Niebur, 2002; Tatler, Baddeley, & Gilchrist, 2005). This kind of guidance can also be learned in the case of domain-specific visual search tasks. For instance, Kundel and & La Follette Jr (1972) explored the difference in eye movement

patterns when expert radiologists versus nonexperts and medical students inspect X-ray images, demonstrating the interplay between top-down and bottom-up factors in guiding attention.

1.4. Our approach

The cognitive studies mentioned above are based on the analysis of small data collected through behavioral lab experiments with children and adults. They usually constrain the kind of visual stimuli presented to participants (synthetic images instead of real-world scenes), so as to make the number of hypotheses limited and explicit. This paper presents a method that allows us to analyze larger datasets and apply the same metrics presented in previous work to inspect human question-asking strategies. To obtain the annotation required to employ these metrics, we leverage powerful computational models and assess their reliability against human participants. In particular, we carry out an in-depth investigation of the informativeness of the questions asked by human participants in a referential visual game, GuessWhat. This task has several similarities with cognitive studies on hypothesis testing and information search using the 20-question game paradigm. Differently from previous work, the two main features of GuessWhat are: (1) the dataset size (tens of thousands of dialogues between human participants); and (2) the use of real-world images. No 20-questions studies, to our knowledge, have presented participants with real objects embedded in real scenes—a more suitable setting to study how humans ask questions in real-world scenarios, and in a very large number of conversations. We adopt the same analytical framework used by previous work in Cognitive Science to investigate players' question-asking strategies. In particular, we analyze the players' strategy effectiveness by computing the EIG of the questions they ask and compare human performance to random and optimal agent simulations. Our analyses validate and confirm findings from previous work, while at the same time allowing us to identify new features of human conversations, such as the unique role of "uninformative" questions (in terms of EIG) when asked in strategic positions. We propose an evaluation framework to assess the effectiveness of complex models based on artificial neural networks to investigate human question-asking strategies in a large number of conversations without the need to hire multitudes of human participants to obtain the necessary annotation.

2. GuessWhat dataset

The GuessWhat visual dialogue game (de Vries et al., 2017) has been proposed within the Natural Language Processing and Computer Vision communities, mostly for engineering purposes. GuessWhat is an asymmetrical game involving two participants who see a real-world image containing several objects. One of the participants (the Oracle) is secretly assigned a target object in the image. The other participant (the Questioner) has to guess the target among the objects appearing in the visual scene by asking the Oracle binary Yes/No questions—extremely similar in structure to the 20-question game used in behavioral studies (see above). The two key subtasks of the Questioner player are thus (1) asking questions; (2) deciding

Fig. 1. A sample game from the GuessWhat dataset. The target object assigned to the Oracle (and that the Questioner has to identify) is highlighted by a green box.

when enough information has been gathered and, hence, stop asking questions and locate the target object in the image among a list of candidates. For each game, it is possible to guess the target only once. Regardless of whether the guess is correct or wrong, it is not possible to continue the dialogue after the guessing action. Participants were instructed that deciding to guess the target would terminate the question-answer exchanges. For this reason, some games finish with a wrong guess about the target object. In total, 85% of the games are successful (target object correctly identified by the Questioner), 8% are unsuccessful (wrong guess about the target), and 7% are unfinished (timeout, disconnection, etc.). Fig. 1 illustrates a sample game from the dataset and Fig. A1 in the online Appendix shows the Questioner interface used during the dataset collection. As mentioned above, the idea of the GuessWhat task can be traced back to the 20-questions game, in which participants try to identify an unknown target object by asking as few yes-or-no questions as possible, either generating the questions from scratch (e.g., Chouinard, Harris, & Maratsos, 2007; Legare, Mills, Souza, Plummer, & Yasskin, 2013; Mosher & Hornsby, 1966; Ruggeri & Lombrozo, 2015; Ruggeri & Feufel, 2015) or selecting them from a list of given alternatives (Nelson et al., 2014; Ruggeri & Lombrozo, 2015). As a peculiarity, GuessWhat uses naturalistic images, thus allowing us to study visual search strategies in complex scenes. The images used in the dataset collection are extracted from the MS-COCO dataset (Lin et al., 2014). By employing Computer Vision tools and models, each image in the dataset is annotated with a set of bounding boxes around objects. The creators of the GuessWhat dataset discarded objects that are too small (area $<$ 500 squared pixels) and only kept images containing 3–20 objects. de Vries et al. (2017) collected the GuessWhat dataset via Amazon Mechanical Turk. The players had to go through a qualification round which consisted of successfully completing 10 games while producing fewer than four mistakes (i.e., wrong guesses), that is, they had to correctly identify the target object in at least six games. After qualification, players were asked to play a batch of 10 successful games. The authors provided bonuses for making fewer mistakes across these 10 games. Moreover, players could report on each other and they were banned after a certain number of reports. Thus, they were encouraged to successfully cooperate and inaccurate players were excluded from the dataset collection procedure. The Questioner was free to ask as many questions as he/she wants, but has to pay a little fee for every new question asked. Overall, the GuessWhat dataset

contains 155K English dialogues about approximately 66K unique images. On average, each dialogue contains 5.2 question-answer pairs. As discussed above, the dataset contains both successful and failed games, that is, games where the Questioner did or did not succeed in identifying the target object at the end of the dialogue (i.e., guessed the wrong object). We only considered successful games (89% of the GuessWhat dialogues are successful). Considering the dataset collection procedure via Amazon Mechanical Turk, unsuccessful dialogues may be very noisy or contain wrong answers from the Oracle (use of automatic bots, lack of attention, etc.).

Most of the work on this task has focused on designing more and more refined computational models to play the role of either the Questioner or the Oracle agent, by implementing findings from NLP and Computer Vision. Different models are usually compared against the Questioner's accuracy in identifying the target object at the end of the dialogue, while little attention has been paid to the efficiency of the questions asked to reach that point. Among the few works that tackle this issue, Zhang et al. (2018); Shuklar et al. (2019); Abbasnejad, Wu, Shi, and van den Hengel (2019) propose to exploit a Deep Reinforcement Learning framework (Sutton & Barto, 2018) for the Questioner model that optimizes reward functions based on the informativeness of the generated questions, while Lee, Heo, and Zhang (2018) use the Oracle agent to directly maximize the information gain of a set of candidate questions. This approach is based on the assumption that the length of the dialogue is a proxy for the efficiency of the questions. However, Mazuecos, Testoni, Bernardi, and Benotti (2020) showed that the questions generated by such models are not efficient, that is, they do not necessarily narrow down the hypothesis space of possible referents. Moreover, Testoni and Bernardi (2021); Testoni, Greco, and Bernardi (2022) showed that State-of-the-Art Guesser models reach a much higher accuracy when receiving human–human exchanges, compared to those generated by the above-mentioned computational models. This strongly suggests that machine-generated dialogues are not as informative as those generated by humans. Interestingly, most previous work in the NLP community assumes that developing conversational agents capable of asking human-like questions means equipping them with optimal asking strategies, that is, including only maximally informative questions. This aspect has never been carefully investigated within this literature. Interestingly, the question has been addressed more extensively in the Cognitive Science community: using a game-setting, Rothe et al. (2018) show that people can accurately evaluate question quality, but have a limited ability to generate maximally informative questions from scratch. Inspired by this work, in our paper, we study whether this finding is confirmed when the visual scene is a real image (hence more complex), with both world knowledge and visual features that could help partition the hypothesis space, as it happens in the GuessWhat game.

3. Overview of the studies

In this paper, we examine for the first time, to our knowledge, the informativeness of the questions asked by human players in a visual-search version of the 20-questions game, the GuessWhat game (de Vries et al., 2017). In particular, we measure the EIG of each question

within the dialogues included in the GuessWhat dataset (described above), to capture the effectiveness of different information search strategies (Study 3). To obtain a lower and upper bound of performance efficiency, we calculated the average EIG related to the performance of simulated Random and Optimal agents.

To measure questions' informativeness, it is necessary to know which objects the players consider as potential target objects when presented with real-world images, that is, what is the *hypothesis space* they entertain. The object annotation released together with the GuessWhat dataset already provides information about the objects presented in each scene. Study 1 aims to assess the reliability of this annotation: given a set of representative scenes, we examined whether human subjects identify the same (number of) objects as in the annotation released with the dataset. This is a necessary condition for us to be able to rely on the released annotation for our analyses in Study 3. We also verified whether some features of the scene impact participants' annotations, and potentially mediate differences between their annotations and those released with the dataset.

To be able to implement our EIG models in Study 3, it is also necessary to understand how players interpret the answers received to their questions, that is, how they narrow down the hypothesis space in response to feedback. Collecting such data for all the scenes included in the GuessWhat dataset would be unfeasible, because of its size. Study 2 aims to check whether the computational model proposed in Testoni et al. (2020) can be effectively adapted to capture the participants' hypothesis space revision process. This model is based on a powerful multimodal Transformer-based architecture (Tan & Bansal, 2019), and it is shown to generate accurate predictions when dealing with language and vision tasks. In our case, we used this model to identify the set of objects in the scene that have the properties described in the question-answer pair under analysis and that, as such, can be possible targets. We then ask human participants to read question-answer pairs extracted from the GuessWhat dataset and flag the objects in the scene that the question targeted. If the model's performance is comparable to participants', we can use the model on the full GuessWhat dataset to automatically determine how the hypothesis space is updated after each question-answer.

4. Study 1: Does participants' initial hypothesis space have the same size as that considered in the released annotations for the GuessWhat game?

4.1. Method

4.1.1. Participants

Participants in Study 1 were 27 university students (17 females and 10 males, $M_{age} = 23.1$ years, $SD = 1.3$ years) recruited from a local university. Written informed consent was obtained from all participants in Study 1 and Study 2. The number of participants for this study was determined by an a priori Student's *t*-Test Power Analysis with a Significance level (alpha) = 0.05, Effect Size (Cohen's *d*) = 0.80, Statistical Power = 80%, which indicates a minimum sample size of 25.5 participants.

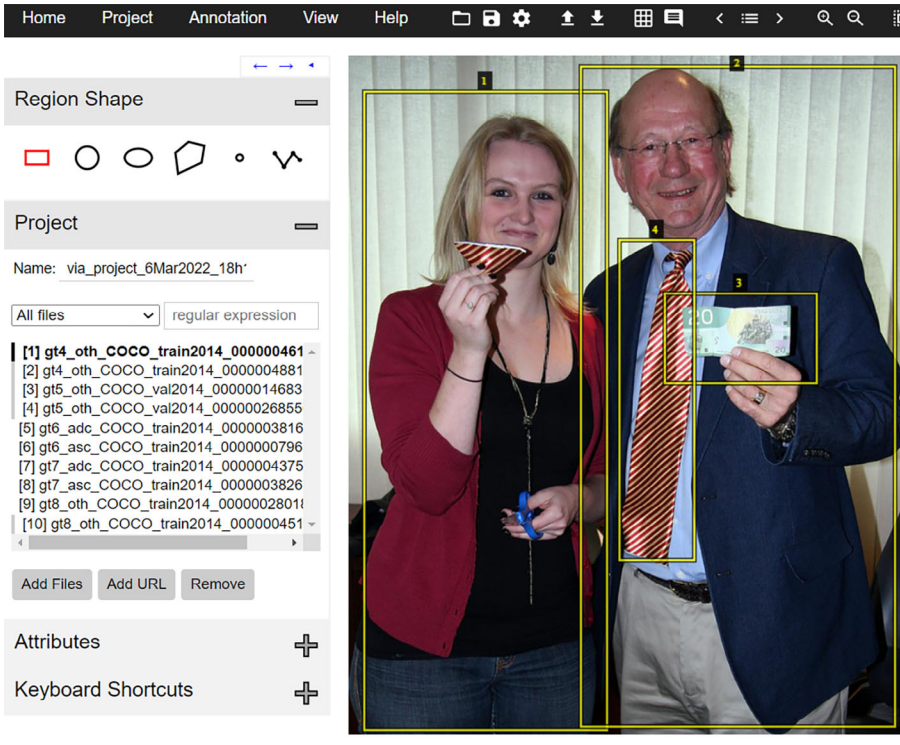

Fig. 2. Annotation software used by participants to draw bounding boxes around possible referents to identify the initial hypothesis space. Source: http://cocodataset.org/#explore?id=461019

4.1.2. Procedure

Participants were tested online using the Zoom platform in groups of 3–5 people. The small size of each group allowed us to supervise participants and make sure they understood the task. At the beginning of the experimental session, we explained participants the GuessWhat game. We clarified that their task was to identify all the possible candidate objects in a visual scene. To collect the data, we adapted the web application proposed in Dutta and Zisserman (2019) (Fig. 2 shows the web interface), which allows participants to draw bounding boxes around candidate objects in the scene. We told them that the number of objects to be identified ranged from a minimum of 4 to a maximum of 20 per scene. During the *training phase*, we showed participants a sample of 10 scenes randomly extracted from the GuessWhat dataset (same scenes for all participants), and we asked them to use the web application to identify all the possible candidate objects in each scene by drawing bounding boxes around the objects (see Fig. 2). Once they had completed the task, we provided them with feedback, that is, we showed them the object annotation extracted from the GuessWhat dataset for the same 10 scenes they had been presented with, and we gave them 5 min (30 s for each scenes) to compare their annotations. This training session was intended to give participants an idea of the level of accuracy and detail we expected from them (e.g., an intuition of how big the objects

had to be for them to be considered). Moreover, it also gave participants the opportunity to familiarize themselves with the web application and ask clarification questions about the task procedure. During the *test session*, participants were randomly assigned a new sample of 10 scenes, different from the ones annotated during the training session, and differing among participants. We did not provide any feedback during the test session.

In total, we presented participants with a subset of 90 scenes from the original dataset, ensuring that each scene would be annotated by three participants. The selected scenes contained 4–8 objects (according to the original dataset's annotation). We also controlled for the percentage of objects of the same category present in each scene: Among the set of 10 scenes, two of them contained only objects of the same category (e.g., two cats, or four cars), two of them did not contain any objects from the same category, and the other six contained intermediate proportions of same/different category objects. Fig. 2 shows the web application interface and an example of the bounding box annotation we asked participants to carry out. The scenes used in this study are available in the shared repository https://osf.io/xywa6/?view_only=6d186c8947ba4c59b860dec1e460e4d5.

4.2. Results

We inspected participants' annotation by calculating, for each participant and for each scene, the absolute difference between the number of boxes in the original dataset's annotations and the number of bounding boxes drawn by the participant.

During the test session, the average absolute difference is 1.63 ($SD = 1.86$, 95% CI [1.40, 1.86]) objects. A closer inspection of participants' individual performance revealed two outliers, who performed much worse than the other participants and likely misunderstood the task goals and/or procedure. We, therefore, decided to exclude their annotations from further analyses. After removing their annotations, we obtained an absolute difference of 1.47 ($SD = 1.42$, 95% CI [1.29, 1.65]). The results obtained examining the relative (signed) difference can be found in the online Appendix. They provide similar insights on the participants' annotation. Furthermore, a manual inspection revealed that the boxes drawn by participants are similar to those released with the dataset not only with respect to their number but also to the objects they identified.

4.2.1. Regression analysis

We conducted a linear regression analysis with absolute difference as a dependent variable, and the between-subjects independent variables expected number of objects (from the original dataset's annotation), round (1–10, representing the order in which the images were annotated), training (before/after the training session, encoded with a value of 0.5 and –0.5, respectively), percentage of objects of the most frequent category in the image (e.g., if the image only contains people, the percentage would be 100%; if the image contains three people, two dogs, and one car, the percentage would be 50%—i.e., three people over six objects in total). We found a significant effect of expected number of objects ($\beta = 0.3746$, $SE = 0.052$, $p < .001$), percentage of objects of the same category ($\beta = -1.6396$, $SE = 0.276$,

$p < .001$), and training ($\beta = 0.6372$, $SE = 0.321$, $p = .047$). We did not find a significant effect of round on absolute difference ($\beta = 0.0063$, $SE = 0.0276$, $p = .818$).

Overall, the results of Study 1 show that the object annotation released together with the GuessWhat dataset is a good approximation of the hypothesis space that human participants consider when presented with a GuessWhat game. In particular, they suggest that participants were more aligned with the dataset's original annotation when the scene contained fewer objects and when most of the objects belonged to the same category.

5. Study 2: Can we automatize the process of hypothesis space revision after feedback?

5.1. Method

5.1.1. Participants

Participants in Study 2 were 24 university students (15 females and 9 males, $M_{age} = 23.34$ years, $SD = 1.44$ years) recruited from a local university.

5.2. Procedure

Participants were tested online using the Zoom platform, in groups of 3– 5 people. The small size of each group allowed us to supervise participants and make sure they understood the task. We first explained to the participants the GuessWhat game. We then illustrated the task they had to carry out: Given a question-answer pair and a scene containing multiple objects extracted from the GuessWhat dataset, participants had to select all the objects (if there were any) that matched the content of the question-answer pair displayed on the screen. In particular, we clarified that they had to identify the set of candidate objects *remaining under consideration* after receiving feedback to a question, that is, all the nonanimals when reading the exchange *"Is it an animal? No."* We showed them some examples of the task they had to perform. We adapted a web application, [2] where participants could click on the ID of the available objects from a list on the right side of the screen (see an illustration of the task in Fig. 3). Participants were free to mark as many objects as they wanted for each dialogue exchange. We used a set of games (i.e., scenes and associated dialogues) extracted from the GuessWhat dataset. We selected a subset of games with 3–6 dialogue exchanges and scenes containing 4–8 objects. Overall, each participant annotated on average 21 games, presented in random order. During the training session, we clarified to participants that, in case they were presented with the same image more than once, the corresponding question-answer pairs were independent of each other and performed by different human participants. We manually inspected the annotation collected for 10 random images and corresponding question-answer pairs and we did not spot any error of this sort. We asked participants to complete the task in 20 min, but all of them finished earlier. In total, 172 games were annotated by three participants each. We compare the performance of the participants of this study to the computational model proposed by Testoni et al. (2020). In particular, this visual question-answer model acts as the GuessWhat Oracle: it receives as input an image, information about the target object, and a question. The output is a Yes or No answer to the input question, based on the target

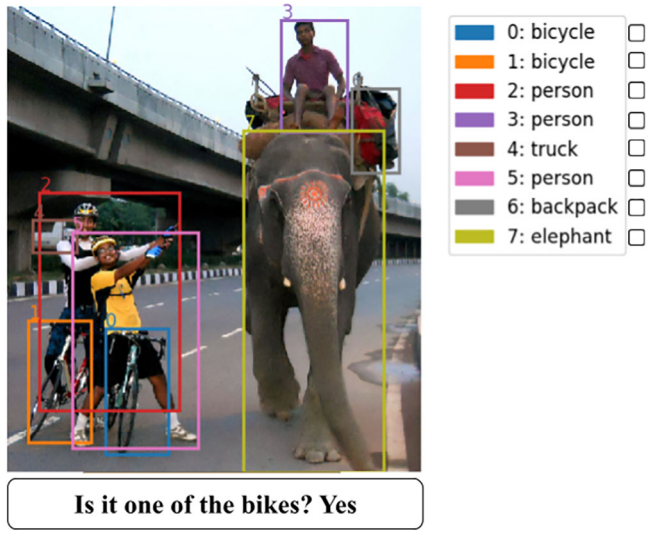

Fig. 3. Example of an image and dialogue exchange displayed to the participants. In this case, they had to select objects 0 and 1 as possible targets.

information and the overall image. For each dialogue turn and question analyzed, we use this model to obtain an answer not just for the target, but for all objects appearing in the image. All the objects that receive the same answer as the target are considered possible targets. In this way, for each question analyzed, all the objects (if there were any) that match the content of the question-answer pair are identified by both the computation model proposed by Testoni et al. (2020) and by the participants of Study 2, allowing us to compare their selection.

5.3. Results

5.3.1. Accuracy

As a sanity check, we computed how often the actual target object was included in the list of possible targets identified by participants at each step. Indeed, by design, the target object for any given dialogue should always be selected and remain under consideration. We refer to this metric as *accuracy*. The question-answer pairs samples from the GuessWhat dataset are not guaranteed to always contain accurate answers. In our studies, given the lack of ground-truth labels, the "accuracy" measure is based on the assumption that human answers are correct. We believe it is fair to make this assumption based on (1) the binary nature of the questions (only yes/no questions); (2) the possibility for the Oracle to answer N/A when the question is not clear; (3) none of the previous work on this dataset has raised issues regarding the quality of human answers; and (4) we manually inspected a large number of question-answer exchanges while designing our studies and we have never spotted any relevant flaw on the answer's side. We found that the accuracy of the participants is 87.4% (95% CI [84.5, 90.2]), meaning that in 87.4% of the dialogue rounds, annotated participants included the actual target in the list of possible targets. The computational model has an accuracy of 86.2% (95% CI [81.5, 90.9]).

5.3.2. Agreement among participants and with the model

We computed the agreement between participants' and model's annotations as to which objects to include as possible targets. We performed this analysis to assess the quality of the model's annotation: if human participants agree with each other as well as with the model, we can assume that the annotation from the model is reliable. To this aim, we check the agreement among participants to interpret the agreement between each participant and the model for the same images and questions. We computed the Krippendorff's alpha agreement coefficient between participants who annotated the same set of games and the model's output. Krippendorff's score generalizes interrater-annotator agreement to an arbitrary number of raters and a variety of different data types. Its value ranges from –1 to 1. As with other agreement measures, it takes into account the random chance of raters to agree on a given observation. We compute the agreement score on each individual question-answer pair: more specifically, we check whether participants decide to mark each object in the image as a possible target. For this reason, the minimum agreement (which depends on the chance level of participants to agree) is bounded by the number of objects in the image. In our experiment, we considered images with at most eight objects, so Krippendorff's score ranges from –0.875 to 1. For every game (annotated by three participants), we computed the score between all possible pairs of participants and the corresponding participant-model agreement. We found that participants agreed with each other (0.61 agreement; 90% CI $[0.57 - 0.65]$) to the same extent as they agreed with the model (0.56 agreement; 90% CI $[0.52 - 0.60]$), t-test: $p = .10$. Considering the range of possible values, both measures can be considered moderate/high agreement. More details about the Krippendorff's score results can be found in the online Appendix. Overall, Study 2 indicates that we can employ the computational model proposed by Testoni et al. (2020) to track how players narrow down the hypothesis space of GuessWhat games based on the feedback received. In Section D in the online Appendix, we present further evidence that Studies 1 and 2 represent a valid estimate of human object identification and hypothesis space updating in relation to the following study.

6. Study 3: Questions' informativeness in the GuessWhat game

6.1. Method

We extracted from the GuessWhat test set all the games successfully solved by human players in less than 10 dialogue exchanges. This resulted in around 19K games overall. For each game, we computed the EIG of each question. In particular, the expected step-wise information gain (see Chin, Payne, Fu, Morrow, & Stine-Morrow, 2015; Nelson, McKenzie, Cottrell, & Sejnowski, 2010; Oaksford & Chater, 1994; Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003) measures the reduction of entropy (Shannon entropy; Shannon, 1948)—that is, the uncertainty as to which hypothesis is correct—upon asking a certain question (see Lindley, 1956). Within this framework, the best questions are the ones that maximize the reduction of entropy, that is, the ones that split the hypothesis space in half (e.g., a question like "is it a person?" at the beginning of the dialogue when the scene contains three people,

two dogs, and one car). Formally, the EIG of each question can be computed by subtracting the expected posterior entropy from the prior entropy:

$$EIG = H_{prior} - H_{posterior}$$

The entropy H represents the uncertainty about which of the candidate hypotheses is true. Its computation is based on the probabilities (p) associated with each of the candidate hypotheses (h). The prior entropy H_{prior} defines the status of uncertainty preceding every action:

$$H_{prior} = -\sum_h p(h) log_2 p(h)$$

The predictive posterior entropy $H_{posterior}$ refers to the predicted uncertainty after the question is asked and the answer is received. The predicted posterior entropy is measured as the sum of the entropies corresponding to each possible future scenario weighted according to the probability of that scenario. Because in our task there are two possible answers to each question (yes/no), $H_{posterior}$ is computed as the sum of:

$$H_{posterior} = p(x_{yes}|X)H(x_{yes}) + p(x_{no}|X)H(x_{no})$$

$p(x_{yes}|X)$ refers to the probability of obtaining a positive answer (assuming that each answer is equally likely). For instance, considering an image containing three dogs and one bear and the question Is *it a bear?*, the probability of obtaining a positive answer is 1 over 4. $H(x_{yes})$ is the entropy observed after receiving a positive answer, computed in the same way as H_{prior} above. We refer to Ruggeri et al. (2017) for more examples of how these entropies are computed.

Building upon the results of Studies 1 and 2, we use the object annotation released in the GuessWhat dataset as the initial hypothesis space, and the model proposed by Testoni et al. (2020) to compute the EIG of each question and update the hypothesis space after each dialogue exchange. In particular, given an image and a question at turn T, the model outputs a Yes/No answer for each object in the image, which represents essential information to compute the EIG. All the objects that receive the same answer as the target are kept as the hypothesis space at turn $T + 1$.

6.2. Simulations

6.2.1. Optimal agent

We modeled the behavior of an optimal agent that, at each step of each game in the dataset, asks the highest EIG a question could possibly achieve, that is, the question that splits the hypothesis space in half (or nearly half, in case of a hypothesis space with an uneven number of hypotheses). Note that we are not interested in the linguistic implementation of this question, but only in its informativeness, that is, in the extent to which it narrows down the hypothesis space currently under consideration. We generally assume that, given the complexity of the scene and the possibility to ask free-form questions, it is always possible to generate such a question. We believe this is a reasonable assumption in spatial contexts because the

Questioner can always resort to literally partitioning the space into quadrants. Another possible strategy relies on listing objects in the hypothesis space with a referring expression for each of them. We considered a version of the optimal agent that selects at each step of the dialogue the question that has the highest EIG, considering the current hypothesis space as defined by the previous human–human dialogue exchange. The Optimal agent is designed to generate the same number of questions as the human Questioner player for each scene in the dataset.

6.2.2. Random agent

We modeled the behavior of a random agent that, at each step of each game in the dataset, asks a question that targets a random proportion of objects in the presented image. At turn T, the random agent generates two random partitions of the N objects left in the hypothesis space at the current turn. Each partition contains a number of objects ranging from 0 to N. The sum of the objects in each partition is equal to N. This simulates the effect of random question-answer exchanges. In our simulation, each random partition of the hypothesis space is equally likely. We compute the EIG using these partitions. The partition containing the target object becomes the hypothesis space at turn $T + 1$. For each dialogue, we generated three random simulations. The random agent is designed to ask exactly the same number of questions of the corresponding human–human dialogue. This choice, in line with previous work that implemented random agents in similar tasks, allows us to isolate the informativeness of the questions asked by fixing the number of questions in the dialogue. Considering only dialogues in which the target was eventually identified, human questioners required on average 4.46 questions. The "stopping rule" described above and the following analyses allow us to draw meaningful conclusions from this experimental setup.

6.2.3. Stopping rule

In the following analyses, we considered two settings when analyzing EIG: We either considered the *full dialogue* (referred to as $Full$ in the Figs.) or we considered the dialogue up to the point where *one* candidate hypothesis was left in the hypothesis space (*truncated dialogue* —referred to as $Truncated$ in the Figs.), that is, excluding all unnecessary questions. In the second case, the stopping point does not correspond to when the human Questioner actually stops, but rather when the hypothesis space is reduced to one single referent. This choice is in line with previous work. For instance, Ruggeri et al. (2016) examine the efficiency of children's *stopping rule* as a potential source of developmental change in question-asking efficiency (see also Chai, Xu, Swaboda, & Ruggeri, 2023). In particular, the authors found that children are significantly more likely than adults to continue their search for information beyond the point at which a single hypothesis remains, and thus to ask questions and select objects associated with zero information gain.

6.3. Results

6.3.1. Participants' EIG versus random and optimal agents

Fig. 4 shows the average EIG when considering the full dialogue ($Full$), revealing that the EIG of questions asked by human players stands between the performance of Random

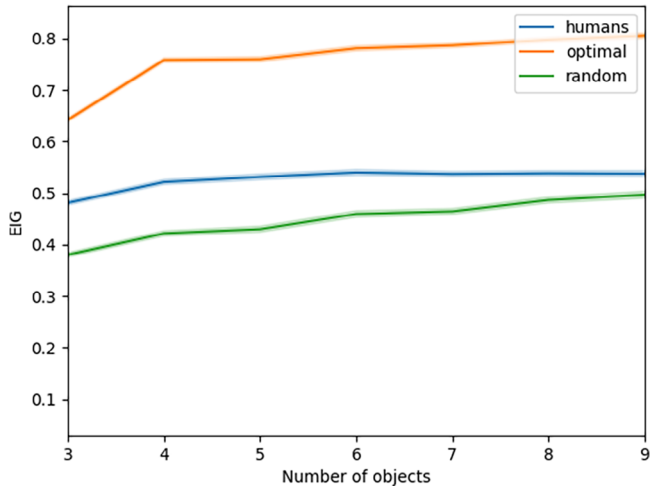

Fig. 4. Average EIG for all the questions asked for images as a function of the number of objects depicted in the image ($Full$). Shaded regions represent 1 SEM in each direction.

and Optimal agents. We conducted a linear regression analysis with EIG as a dependent variable, and the between-subjects independent variables number of objects and model (levels: humans, optimal, random). We found a significant effect of number of objects ($\beta = 0.0061$, $SE = 0.001$, $p < .001$) and model (optimal: $\beta = 0.2503$, $SE = 0.002$, $p < .001$, random: $\beta = -0.0442$, $SE = 0.002$, $p < .001$). We can thus conclude that the average EIG is higher when the scene contains more object and the model simulations (optimal and random agents) have a significant effect in opposite directions compared to human performance.

When considering the truncated dialogues (Fig. 5), we can see instead that the Random agent generally outperforms human performance. A linear regression analysis, identical to that ran above, indicates again a significant effect of number of objects ($\beta = 0.0019$, $SE = 0.001$, $p < .001$) and model (optimal: $\beta = 0.3187$, $SE = 0.002$, $p < .001$, random: $\beta = 0.0784$, $SE = 0.002$, $p < .001$). In this case, both random and optimal agents reveal a significantly higher EIG compared to human performance. This discrepancy between full and truncated dialogues suggests that the advantage of human players compared to Random agents when considering the full dialogues (Fig. 4) was due to a more efficient stopping criterion. In the following, we explore this hypothesis by examining the impact of *"uninformative"* questions.

6.3.2. Participants' versus random agents' "uninformative" questions

We define a question to be "uninformative" if its EIG equals 0. In the following, we provide evidence that this term may be misleading, and we hypothesize that questions of this sort are strategically used by players during the dialogue. Fig. 6 shows that, if we consider the full dialogues, the Random agents always generate more uninformative questions than human players (McNemar's1947 test, $p < .01$ for all comparisons given the same number of objects), especially when there are few objects in the scene. However, when considering

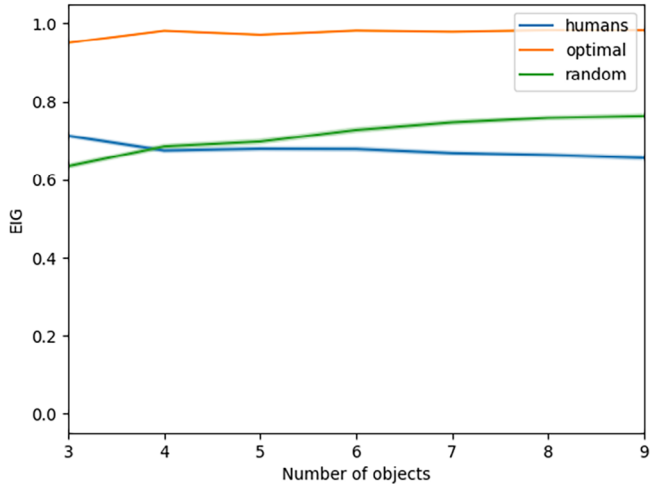

Fig. 5. Average EIG for all the questions asked for images as a function of the number of objects depicted in the image (*Truncated*). Shaded regions represent 1 SEM in each direction.

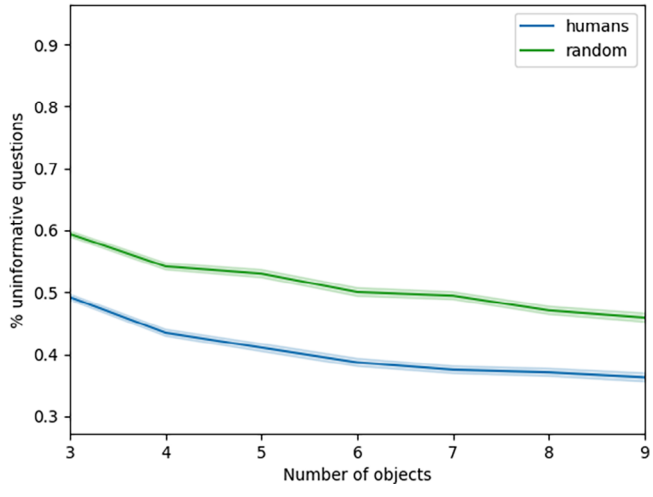

Fig. 6. Percentage of uninformative questions against the number of objects in the scene (*Full*). Shaded regions represent 1 SEM in each direction.

truncated dialogues (see Fig. 7), we observe a different pattern: Humans ask on average a similar percentage of uninformative questions (around 25%), regardless of the number of objects in the scene. The random agent, instead, shows a monotonical decrease in the percentage of uninformative questions with more objects in the scene. According to a z-test for proportions, the difference between human and random proportion of uninformative questions is not statistically significant when the image contains four or five objects ($p = .705$ and $p = .031$, respectively). In all the other cases, the difference is statically significant

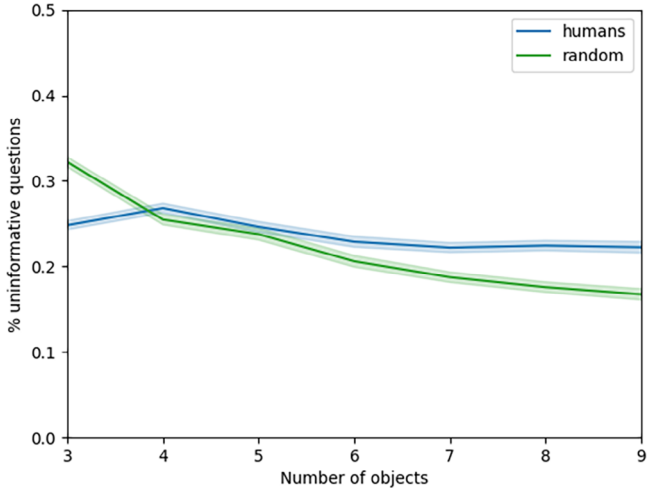

Fig. 7. Percentage of uninformative questions against the number of objects in the scene (*Truncated*). Shaded regions represent 1 SEM in each direction.

($p < .01$), with humans significantly outperforming the random agent in scenes with six or more objects. This result suggests that uninformative questions may play a peculiar role in the human-dialogue strategy, regardless of the number of objects in the scene, while a random strategy is solely influenced by this number.

We examined the average number of additional questions asked *after* the stopping point. We found that humans, on average, ask one additional uninformative question after they reduced the hypothesis space to one candidate (average = 1.13 questions, 90% CI [1.11, 1.15]). We found that this question generally summarizes the information collected throughout the dialogue, thus we refer to it as a confirmation question. Indeed, it receives a positive answer 70% of the time. Fig. 8 shows two examples of human–human dialogues containing a confirmation (uninformative) question as the final question. The Random agents, instead, ask on average two additional uninformative questions after the stopping point (average = 2.52, 90% CI [2.43, 2.61]). The average number of extra uninformative questions is significantly different for humans and the random agent, according to a *t*-test with a significance level of 0.05 ($t = –47.29$, $p < .01$). This result explains the numbers reported above: When considering the full dialogues, the Random agent is penalized because, by design, for any given image, it has to ask the same number of questions that human players asked and, thus, it is likely to ask more uninformative questions after the hypothesis space is reduced to one object. On the contrary, while human players may take longer to identify the target, they usually ask only one additional question that does not bring further information about the target. We hypothesize that, given the complexity of the scenes in this task, humans may ask one additional "uninformative" confirmatory question at the end of the dialogue to double-check their hypothesis before guessing. A similar phenomenon was found also in referential games with much simpler scenes (Ruggeri et al., 2016), pointing to the challenging nature of establishing stopping

Is it a human? No
Is it a piece of clothing/apparel? No
Is it yellow/almost yellow? Yes
Is it in the left half of the image? Yes

Is it text? No
Is it a clock? Yes
Does it have TRAIN written right above it? No
Does it have STATION written right above it? Yes

Fig. 8. Two examples of dialogues containing a confirmation question at the end of the dialogue. Though uninformative, dialogues often contain this kind of question at the end.

rules. In the following, we further investigate the role of "uninformative" questions asked in different dialogue turns.

6.3.3. EIG throughout the dialogue

We explored how the questions' EIG changes throughout the dialogue. In the following, we consider images with 4–8 candidate objects. We found that controlling for the number of objects in the image leads to the same results and patterns discussed in the following. We observe interesting phenomena when comparing human and random performance across dialogue exchanges. Fig. 9 shows that in the first exchange (turn 0), the average EIG of the questions asked by the Random agents is significantly higher than the one observed in human players (paired t-test, $p < .001$). Starting from the second dialogue exchange, human EIG always significantly outperforms the random agent simulation ($p < .001$). We observe a similar pattern when considering the truncated dialogues, see Fig. 10, with the random agent simulations outperforming human players in terms of EIG at the beginning of the dialogue (t-test, $p < .001$). Note that in Fig. 9 the optimal agent's average performance is dropping

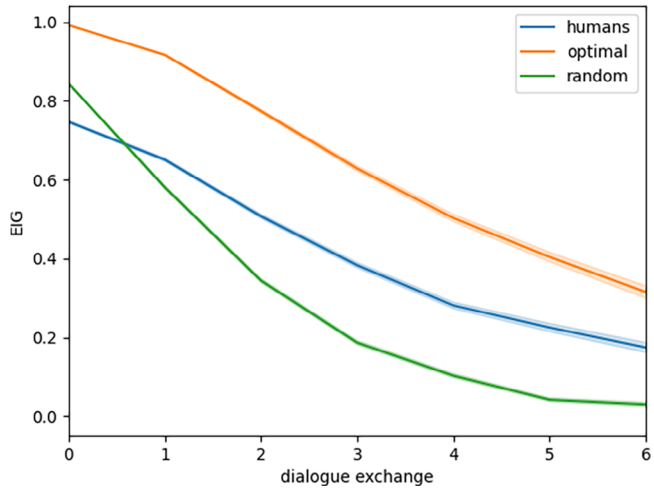

Fig. 9. EIG across dialogue exchanges (x-axis) for images with 4–8 objects ($Full$). Shaded regions represent 1 SEM in each direction.

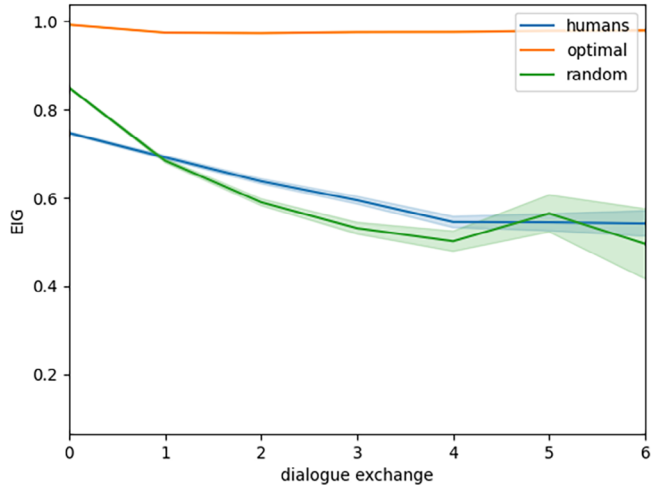

Fig. 10. EIG across dialogue exchanges (x-axis) for images with 4–8 objects ($Truncated$). Shaded regions represent 1 SEM in each direction.

across the full exchanges because the model asks the same number of questions as the human Questioner for a given scene.

To shed further light on this result, we examined the percentage of uninformative and optimal (i.e., maximally informative in the considered hypothesis space) questions during the dialogue. We distinguished between games with fewer objects (4–5 objects, Fig. 11) and games with more objects (6–8 objects, Fig. 12). Nonoptimal but informative questions are not shown in the image, but they represent the remaining proportion of questions for each

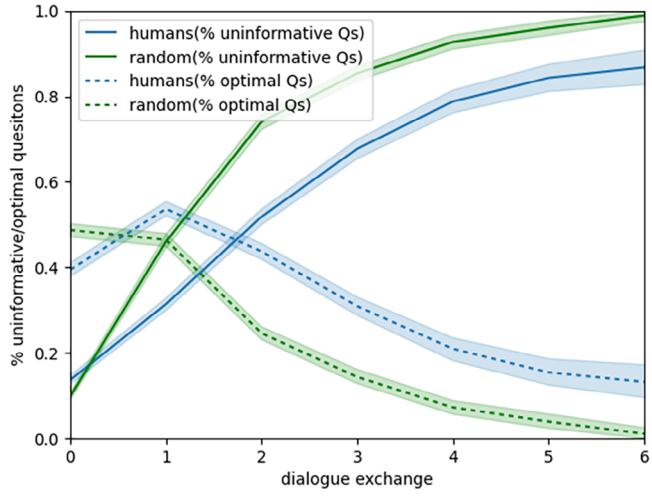

Fig. 11. Percentage of uninformative and optimal questions in games with few objects (4–5 objects) against dialogue exchange (*x*-axis). Shaded regions show 95% confidence intervals (95% CIs). (Full)

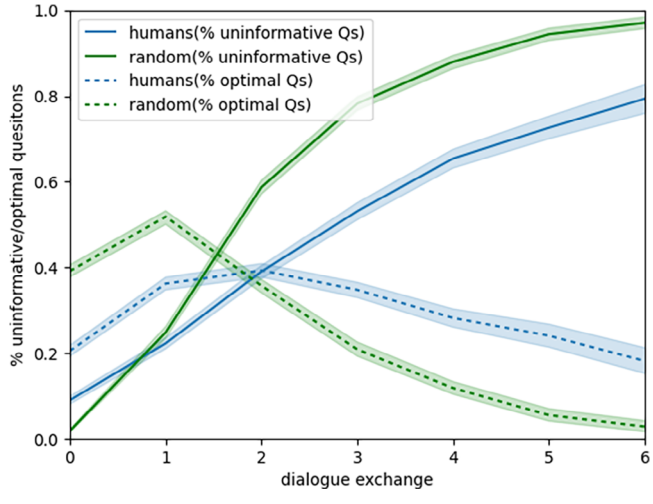

Fig. 12. Percentage of uninformative and optimal questions in games with many objects (6–8 objects) against dialogue exchange (*x*-axis). Shaded regions show 95% confidence intervals (95% CIs). (Full)

exchange. We can observe some interesting phenomena by comparing the performance of the random agent with the one of human players. All the differences discussed in the following are significant according to a McNemar test (McNemar, 1947) with a significance level of 0.05. In general, humans struggle to generate optimal questions, in particular at the beginning of dialogues referring to scenes containing many objects. In this case, in fact, the random agent outputs significantly more optimal questions than human players. Not only the random agent generates more optimal questions at the beginning of the dialogue, but it also generates

Is it an animal? Is it a red plane? Is it food?

Fig. 13. Three examples of uninformative questions asked at the beginning of the dialogue. We conjecture they serve as a way to establish a common ground with the interlocutor.

fewer uninformative questions than human players. In the next section, we investigate the role of these questions and formulate a hypothesis about this surprising result.

6.3.4. The role of "Uninformative" questions at the beginning of the dialogue

Focusing on the very first turn of the dialogue in scenes with many objects, we find that humans ask significantly more "uninformative" questions than the random agent (9.06% vs. 1.93%, $p < .001$). The reason why human players ask uninformative questions, however, is not clear. In the following, we wonder whether these questions are indeed strictly uninformative or whether their unexpectedly high percentage reveals that they play a pragmatic role in the dialogue strategy.

We define as *common-ground-establishing* questions those questions asked at the beginning of the dialogue that display two peculiar features: being "uninformative," as measured by their EIG (i.e., they do not narrow down the set of possible targets), and receiving a positive answer. In this way, we make sure that the question applies to all referents in the scene. We checked the percentage of such questions asked at the beginning of the dialogue that receive positive feedback and compare them with the other questions in the same position. We found that the majority of the "uninformative" questions asked at the beginning of the dialogue receive a positive response (67.0% of the time), while only 50% of "informative" questions asked by human players in the same position receive positive answers. Nonoptimal and optimal questions (medium to high EIG) in the same position get predominantly negative answers (52.1% of negative answers). If we limit our analysis to images with 4–8 objects for consistency with Studies 1 and 2, we find that "uninformative" questions get positive answers 61.3% of the time, while only 47.7% of the questions asked by human participants (with EIG > 0) in the same position receive positive feedback. As mentioned above, all the differences are significant according to a McNemar test.

Fig. 13 shows three examples of questions that are asked in the first turn of the dialogue, are uninformative ($EIG = 0$), and receive a positive answer. Note that participants went through a training phase during which they learned which objects in the image are potential targets. Nevertheless, in the examples, the questioner looks for confirmation of such training, and asks a question confirming that the target is one of the objects in the image— for example, "is it an animal?" when all the candidate objects (according to the dataset annotation, which

participants aligned to during the training phase) are animals, or "Is it a red plane?" when all the candidate objects are red planes.

7. General Discussion

Solving a task by asking questions requires a plethora of different skills, and modeling such effort is not trivial. In this paper, we analyzed—to our knowledge, for the first time—humans' question-generation performance in the context of a large-scale visual search task. Generating informative questions is generally a very sophisticated competence, as it requires verbal knowledge and, potentially, a rich vocabulary, categorization skills, and taps on previous experience. For example, to generate questions that target groups of objects, one needs to be able to identify features that can be used to group objects into different categories, categorize objects correctly according to those features, and label those categories to formulate and then utter the appropriate question (see Ruggeri et al., 2017). Asking questions is even more complex in real-world visual search tasks, as it additionally taps onto humans' visual skills and perception (e.g., capacity to discern colors, to isolate individual objects from the background, etc.), potentially inducing an even greater variability across individuals. It is thus particularly important that the agents at the two ends of the communication channel agree with each other and establish a common ground (e.g., consider the same hypothesis space).

In this work, we propose to take advantage of computational tools to obtain the annotation needed to apply metrics and evaluation practices from cognitive studies on a large-scale. We take the GuessWhat visual dialogue referential task as a test-bed. GuessWhat is a task inspired by the 20-question setting designed within the Natural Language Processing and Computer Vision communities to develop multimodal goal-oriented conversational agents. In Guesswhat, the conversation between the Questioner and the Oracle/Answerer builds upon complex real-world images. The dataset contains tens of thousands of human–human dialogues on as many different images. We investigate the strategy that human players follow when playing the game as the Questioner agent. In our first study (Study 1), we checked whether the automatic object-level annotation associated with each image in the GuessWhat dataset represents a good approximation of the initial hypothesis space of human players. We found that participants identify a number of objects similar to the one reported in the dataset. This annotation is particularly reliable for scenes with few objects and with objects belonging to the same category. We could thus trust the automatic object-level annotation as a good proxy for the initial hypothesis space human players entertain in the GuessWhat dataset.

In Study 2, we concluded that a computational model is as accurate as human participants in tracking how the hypothesis space changes after each dialogue exchange. Moreover, we found that humans agree with each other as much as they do with the computational model. They differ with respect to the accuracy in processing positive and negative feedback coming from the Oracle, as highlighted in the Supplementary Material (Section C). Overall, our results demonstrate that we can effectively automatize the process of hypothesis space revision after feedback. This is a crucial step, as it allows us to borrow metrics from a body of cognitive works and apply them to a much larger dataset. Studies 1 and 2 provided us with the *ingredients* to compute the EIG of the questions asked by human players in the dialogues

included in the GuessWhat dataset. Following previous work, to better inspect the effectiveness of the strategy followed by human players, we compared their performance against simulations of optimal and random agents. First of all, it is quite comforting to report that human players do perform, on average, better than a random model. However, their question-asking performance when playing visual referential games is far from optimal. In particular, when considering truncated dialogues (i.e., the part of the dialogue *before* the hypothesis space is narrowed down to only one candidate hypothesis), we found that human players' questions are sometimes less informative than those produced by a random simulated agent, especially in scenes containing many objects.

Our in-depth analysis allows us to unveil the crucial but underinvestigated role of "uninformative questions" throughout the dialogue exchanges. On the one hand, we found that human players frequently asked uninformative questions at the beginning of the dialogue. We hypothesize that players establish a common ground with the interlocutor by asking questions that target specific features of the scene (irrespective of their utility for identifying the referent), relying prominently on cognitive or perceptual salience. These questions are meant to agree on *what is* the hypothesis space. On the other hand, toward the end of the dialogue, when the hypothesis space is reduced to only one candidate object, humans tend to ask only one confirmation question, that is, a question that summarizes the information collected through the dialogue, describing the target and, therefore, most often obtains a positive answer. It thus arises the question of whether conversational agents trained to solve this task should learn these suboptimal features characterizing human dialogues, as well. These features could make the text generated by computational models sound more human-like, and they may also represent useful pragmatic "tricks" to make the conversation more successful.

Our proposed method demonstrates some of the benefits that can be obtained from the application of cognitive science techniques and metrics on a large scale, unveiling phenomena characterizing human language that can hardly be inspected on small datasets. Our results open new questions about the interaction between players in cooperative visual search tasks. It is worth noting that in the GuessWhat dataset games, participants were paired through a chat interface and could not see each other. Would a dialogue between two people playing the game face-to-face display different features? How does trust in the partner, perspective taking competences, previous experience, and visual skills impact the interaction—its success, efficiency, and effectiveness? Moreover, it would be interesting to further explore our findings from a development perspective to unveil how children and adults differ with respect to the visual search dialogue strategies we analyzed in this work. Previous work in this direction found a clear developmental trajectory in the effectiveness of children's question-asking strategies. These studies, however, made use of simpler, artificial visual inputs and it is not clear whether the same pattern would hold when dealing with complex visual stimuli as in GuessWhat.

Finally, it would be interesting to understand how the dynamic change of the hypothesis space during the dialogue relates to the actual eye fixations of the Questioner player, a relevant aspect of visual search that we did not investigate. Do participants keep on scanning the objects after they have been excluded from previous dialogue exchanges, or do they only focus on the objects still under consideration? Eye fixation patterns represent an effective source to investigate the high-level cognitive skills required to solve the Guesswhat task. This analysis may also unveil whether the inability of players to ask optimal questions highlighted

in our work comes from a partial or inadequate scanning of the visual scene. If this is the case, eye fixations may help to explore the perceptual salience features that guide human players when asking questions.

8. Conclusion

In this paper, we propose to bring together efforts and techniques from different disciplines (in particular, Computer Science and Cognitive Science) to investigate human behavior and language use on large datasets of human conversations. We focus on a visual search dialogue task as a case study to show how to leverage computational models in order to obtain the annotation needed to apply Cognitive Science metrics and evaluation techniques on larger datasets. We also highlight the benefits of this approach, which allows us to corroborate and extend findings from previous work restricted to smaller and less natural settings. In our paper, we focus on GuessWhat, a visual search dialogue task inspired by the 20-question game setting. Visual search tasks guided by the interaction with one or multiple interlocutors are the backbone of many daily activities. In their complexity, they involve a plethora of high-level cognitive skills and have attracted the attention of research communities across different fields. On the one hand, cognitive scientists in their search experiments generally make strong simplifications about the visual stimuli presented, and are usually limited to a few data observations. On the other hand, other research communities mainly aim at building huge datasets of human–human conversations to feed data-hungry artificial neural network models, without paying much attention to the psychological plausibility of the paradigms used, or to the features of the collected data. Our work aims at filling this gap by providing a comprehensive set of studies to unveil the question-asking strategy of a large set of human players playing a complex referential visual dialogue task. We demonstrate the effectiveness of our method which consists of using computational tools to automatically code and capture how the hypothesis space changes throughout the dialogue in complex real-world scenes, paving the way for future research in this direction—also from a developmental perspective. Moreover, our analyses highlight the challenges behind modeling pragmatic phenomena that characterize human conversations, emphasizing, in particular, the role of "uninformative" questions as measured by their EIG. Our work poses new questions about the development of intelligent conversational systems that can handle the complexity of human language. We emphasize the importance of developing conversational agents that do not necessarily ask "optimal" questions, but instead are capable of mastering those "pragmatic tricks" (such as asking "uninformative" questions at the proper round of the dialogue) that make human communication successful. We leave this as open discussion, as it heavily depends on the application domain, the target public, and the required outcome.

Acknowledgments

We are grateful to Oana Stanciu and Sandro Pezzelle for providing feedback and support on preliminary versions of this work. We would like to express our sincere gratitude to the

anonymous Reviewers and the Executive Editor for their valuable comments and insightful discussion. Alberto Testoni started this work as a PhD student at the University of Trento, affiliated to the Department of Information Engineering and Computer Science (DISI). Alberto Testoni is currently funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 819455, PI R. Fernández).

Notes

1 https://chat.openai.com/chat
2 https://www.makesense.ai/

References

Abbasnejad, E., Wu, Q., Shi, J., & van den Hengel, A. (2019). What's to know? Uncertainty as a guide to asking goal-oriented questions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4155–4164).

Belke, E., & Meyer, A. S. (2002). Tracking the time course of multidimensional stimulus discrimination: Analyses of viewing patterns and processing times during "same"-"different" decisions. *European Journal of Cognitive Psychology*, *14*(2), 237–266.

Chai, K.-X., Xu, F., Swaboda, N., & Ruggeri, A. (2023). Preschoolers' information search strategies: Inefficient but adaptive. *Frontiers in Psychology*, *13*, 1080755.

Chan, L. K., & Hayward, W. G. (2013). Visual search. *Wiley Interdisciplinary Reviews: Cognitive Science*, *4*(4), 415–429.

Chin, J., Payne, B. R., Fu, W.-T., Morrow, D. G., & Stine-Morrow, E. A. (2015). Information foraging across the life span: Search and switch in unknown patches. *Topics in Cognitive Science*, *7*(3), 428–450.

Chouinard, M. M., Harris, P. L., & Maratsos, M. P. (2007). Children's questions: A mechanism for cognitive development. *Monographs of the Society for Research in Child Development*, i–129.

de Vries, H., Strub, F., Chandar, S., Pietquin, O., Larochelle, H., & Courville, A. C. (2017). Guesswhat?! Visual object discovery through multi-modal dialogue. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21–26, 2017* (pp. 4466–4475). IEEE Computer Society.

Dutta, A., & Zisserman, A. (2019). The VIA annotation software for images, audio and video. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, New York, NY, USA. ACM.

Eckstein, M. P. (2011). Visual search: A retrospective. *Journal of Vision*, *11*(5), 14.

Eikmeyer, H.-J., & Ahlsén, E. (1996). The cognitive process of referring to an object: A comparative study of German and Swedish. In *Proceedings of the 16th Scandinavian Conference on Linguistics, Turku, Finland*.

Gatt, A., van Gompel, R. P., van Deemter, K., & Krahmer, E. (2013). Are we Bayesian referring expression generators. In *Proceedings of the Workshop on Production of Referring Expressions: Bridging the Gap between Computational and Cognitive Approaches to Reference (PRE-CogSci'13)*. Berlin.

Gurari, D., Li, Q., Stangl, A. J., Guo, A., Lin, C., Grauman, K., Luo, J., & Bigham, J. P. (2018). Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3608–3617).

Hayhoe, M., & Ballard, D. (2005). Eye movements in natural behavior. *Trends in Cognitive Sciences*, *9*(4), 188–194.

Hayhoe, M. M., Shrivastava, A., Mruczek, R., & Pelz, J. B. (2003). Visual memory and motor planning in a natural task. *Journal of Vision*, *3*(1), 6.

Herwig, J. E. (1982). Effects of age, stimuli, and category recognition factors in children's inquiry behavior. *Journal of Experimental Child Psychology*, *33*(2), 196–206.

Kundel, H. L., & La Follette Jr, P. S. (1972). Visual search patterns and experience with radiological images. *Radiology*, *103*(3), 523–528.

Land, M., Mennie, N., & Rusted, J. (1999). The roles of vision and eye movements in the control of activities of daily living. *Perception*, *28*(11), 1311–1328.

Lee, S.-W., Heo, Y.-J., & Zhang, B.-T. (2018). Answerer in questioner's mind: Information theoretic approach to goal-oriented visual dialog. In *Advances in Neural Information Processing Systems* (pp. 2579–2589).

Legare, C. H., Mills, C. M., Souza, A. L., Plummer, L. E., & Yasskin, R. (2013). The use of questions as problem-solving strategies during early childhood. *Journal of Experimental Child Psychology*, *114*(1), 63–76.

Lin, T., Maire, M., Belongie, S. J., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In D. J. Fleet, T. Pajdla, B. Schiele & T. Tuytelaars (Eds.), *Computer Vision - ECCV 2014-13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science* (pp. 740–755). Springer.

Lindley, D. V. (1956). On a measure of the information provided by an experiment. *Annals of Mathematical Statistics*, *27*(4), 986–1005.

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2023). Dissociating language and thought in large language models: A cognitive perspective. *arXiv preprint arXiv:2301.06627*.

Mazuecos, M., Testoni, A., Bernardi, R., & Benotti, L. (2020). On the role of effective and referring questions in GuessWhat?! In *Proceedings of the First Workshop on Advances in Language and Vision Research* (pp. 19–25). Association for Computational Linguistics.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, *12*(2), 153–157.

Meder, B., & Nelson, J. D. (2012). Information search with situation-specific reward functions. *Judgment and Decision Making*, *7*(2), 119–148.

Nelson, J. D. (2005). Finding useful questions: On Bayesian diagnosticity, probability, impact, and information gain. *Psychological Review*, *112*(4), 979.

Nelson, J. D., Divjak, B., Gudmundsdottir, G., Martignon, L. F., & Meder, B. (2014). Children's sequential information search is sensitive to environmental probabilities. *Cognition*, *130*(1), 74–80.

Mosher, F. A., & Hornsby, J. R. (1966). On asking questions. In J. S. Bruner, R. R. Oliver & P. M. Greenfield, et al. (Eds.), *Studies in cognitive growth* (pp. 86–102). New York, NY: Wiley.

Nelson, J. D., McKenzie, C. R., Cottrell, G. W., & Sejnowski, T. J. (2010). Experience matters: Information acquisition optimizes probability gain. *Psychological Science*, *21*(7), 960–969.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, *101*(4), 608.

Parkhurst, D., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, *42*(1), 107–123.

Rothe, A., Lake, B. M., & Gureckis, T. M. (2018). Do people ask good questions? *Computational Brain & Behavior*, *1*(1), 69–89.

Ruggeri, A., & Feufel, M. (2015). How basic-level objects facilitate question-asking in a categorization task. *Frontiers in Psychology*, *6*, 918.

Ruggeri, A., & Lombrozo, T. (2015). Children adapt their questions to achieve efficient search. *Cognition*, *143*, 203–216.

Ruggeri, A., Lombrozo, T., Griffiths, T. L., & Xu, F. (2016). Sources of developmental change in the efficiency of information search. *Developmental Psychology*, *52*(12), 2159.

Ruggeri, A., Sim, Z. L., & Xu, F. (2017). "Why is toma late to school again?" Preschoolers identify the most informative questions. *Developmental Psychology*, *53*(9), 1620.

Ruggeri, A., Walker, C. M., Lombrozo, T., & Gopnik, A. (2021). How to help young children ask better questions? *Frontiers in Psychology*, *11*, 586819.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379–423.

Shuklar, P., Elmadjian, C., Sharan, R., Kulkarni, V., Wang, W. Y., & Turk, M. (2019). What should I ask? Using conversationally informative rewards for goal-oriented visual dialogue. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 6442–6451).

Singh, S., & Beniwal, H. (2022). A survey on near-human conversational agents. *Journal of King Saud University-Computer and Information Sciences*, *34*(10), 8852–8866.

Sordoni, A., Bengio, Y., Vahabi, H., Lioma, C., Grue Simonsen, J., & Nie, J.-Y. (2015). A hierarchical recurrent encoder-decoder for generative context-aware query suggestion. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management* (pp. 553–562).

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E.-J., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science*, *27*(3), 453–489.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT Press.

Tan, H., & Bansal, M. (2019). Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 5100–5111).

Tatler, B. W., Baddeley, R. J., & Gilchrist, I. D. (2005). Visual correlates of fixation selection: Effects of scale and time. *Vision Research*, *45*(5), 643–659.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*(4), 629–640.

Testoni, A., & Bernardi, R. (2021). The interplay of task success and dialogue quality: An in-depth evaluation in task-oriented visual dialogues. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (pp. 2071–2082). Association for Computational Linguistics.

Testoni, A., Greco, C., & Bernardi, R. (2022). Artificial intelligence models do not ground negation, humans do. Guesswhat?! Dialogues as a case study. *Frontiers in Big Data*, *4*. https://doi.org/10.3389/fdata.2021.736709

Testoni, A., Greco, C., Bianchi, T., Mazuecos, M., Marcante, A., Benotti, L., & Bernardi, R. (2020). They are not all alike: Answering different spatial questions requires different grounding strategies. In *Proceedings of the Third International Workshop on Spatial Language Understanding* (pp. 29–38). Association for Computational Linguistics.

Treisman, A. (1982). Perceptual grouping and attention in visual search for features and for objects. *Journal of Experimental Psychology: Human Perception and Performance*, *8*(2), 194.

Yarbus, A. L. (1967). Eye movements during fixation on stationary objects. In *Eye movements and vision* (pp. 103–127). Boston, MA: Springer.

Zhang, J., Wu, Q., Shen, C., Zhang, J., Lu, J., & van den Hengel, A. (2018). Goal-oriented visual question generation via intermediate rewards. In *Proceedings of the European Conference of Computer Vision (ECCV)* (pp. 186–201).

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Fig. A1. Questioner interface used in the GuessWhat dataset collection.

Fig. B1. Absolute difference (*y*-axis) of each participant (reported on the *x*-axis), before and after the training session (red and green, respectively).