Referring Effectively to a Group of Objects

Emilia Ellsiepen (emilia@lst.uni-saarland.de), Alexandra Mayn, Natalia Bila & Vera Demberg
Department of Language Science and Technology, Department of Computer Science
Saarland Informatics Campus, Saarland University

Abstract

In communication, we often need to refer to multiple objects. We have several possibilities: enumerating the objects one by one or referring to a shared feature. This is the first study to systematically compare different ways of referring to groups of objects in terms of response times, accuracy, and memory encoding. In a within-subjects design, participants listened to instructions which either used a shared feature or enumerated coordinates of individual objects. We find that there is no one way of referring that is universally more effective, but the optimal way of referring depends on the type of common feature and the number of objects. When the shared feature is prominent, such as color, and multiple objects need to be identified, referring to it results in shorter identification times and fewer errors. In contrast, when the shared feature is less salient, such as pattern, referring by location tends to be better.

Keywords: referring expressions; psycholinguistics; visual search

Introduction

When we communicate, we often need to point out objects to our interlocutor. The way people refer has been a subject of investigation in pragmatics and psycholinguistics. When referring to multiple objects as opposed to just one, there are multiple possibilities. Consider the array of objects in Figure 1. To refer to the same set of objects, we can enumerate their locations ("A1, A3, C2 and D4"), or refer to their common feature ("the blue objects"). Enumerating locations has the advantage that the listener knows how many objects they are supposed to attend to and that they do not have to engage in exhaustive visual search. A common feature, on the other hand, can be expressed more efficiently, i.e., with a shorter utterance. But which type of reference is more effective in the end? How quickly and accurately does the listener establish reference to all members of the group for the two types of reference? And what role does the size of the subgroup and the kind of common feature play in this task?

In this paper, we present an experiment which investigates whether one type of reference is generally more optimal compared to the other in terms of identification time, identification accuracy, and memory encoding for a display containing 16 abstract objects of similar size. By varying the number of objects that needs to be picked out, we furthermore investigate whether there exists a trade-off between the length and the specificity of the utterance. Finally, we contrast two common features of different saliency, namely color and pattern.

This work has implications for three different research areas: psycholinguistics, where previous literature on reference has mainly focused on production; visual search, where existing work on groups of objects is sparse; and human-computer interaction, where guidelines on which reference types are more effective in which contexts would benefit the development of successful interfaces, e.g., in a monitoring task with multiple entities arranged in a grid structure such as our example.

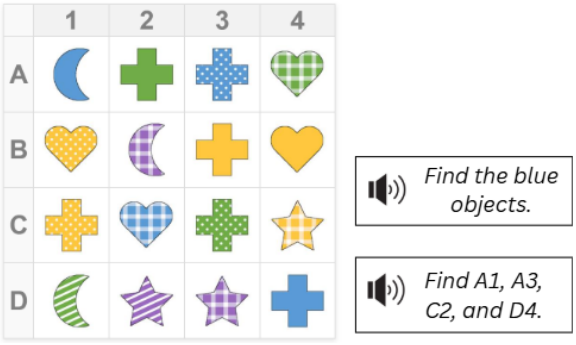


Figure 1: Example trial in both conditions.

Previous work

Psycholinguistic research has looked at different types of reference mainly from the production perspective and for single objects. Speakers often engage in overspecification, i.e., in a given referential context, more attributes are produced than strictly necessary to identify the referent unambiguously. Research on this topic has revealed that overspecification occurs more often for more prominent features, like color adjectives (Pechmann, 1989), but it also depends on the number of objects to be described and the complexity of the scene (Elsner et al., 2018; Koolen et al., 2011). Much less work has been conducted on the ease of *comprehending* different types of referring expressions. Here, existing work has mainly addressed the question of whether overspecification is beneficial, neutral, or detrimental to the process of establishing reference to a single referent. In an eye-tracking study, Rubio-Fernandez (2021) showed that overspecified color adjectives speeded up identification in terms of fixation latency and reaction time in comparison to the shorter utterance in similar visual contexts that triggered the production of overinformative adjectives. From this they argue that informativity is not the only relevant

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

dimension, but visual discriminatory value is also important. Arts et al. (2011) found that overspecification can be beneficial, as evidenced by reduced search times, especially if the referring expression includes location information. Similarly, Fukumura and Carminati (2022) found a benefit for the highly salient attribute color, but not for pattern in overspecified referring expressions. This suggests that ease of identification heavily relies on the type of feature used to refer. In contrast, other studies have found detrimental effects of overspecification (e.g., Engelhardt et al., 2011).

While our work does not address overspecification in the sense of an additional adjective, there is a clear connection with regard to the debate on whether overspecification does or does not violate the maxim of quantity (Grice, 1975), if interpreted strictly in terms of length: When reference by a common feature is possible, it leads to shorter descriptions in the multi-object scenario in a similar way that minimal descriptions are shorter than overspecified ones. Assuming that violations of the maxim of quantity in this sense would lead to difficulties on the listener's part, the expectation would therefore be that the shortest utterance that uniquely identifies the target objects should lead to optimal identification, i.e., reference by a common feature should be better for multiple objects, irrespective of the type of common feature. Alternatively, in line with the visual efficiency hypothesis (Rubio-Fernandez, 2021), longer descriptions might be more effective if they facilitate the search process.

Most of the studies mentioned above use very simple displays, e.g., containing only 4 objects (Arts et al., 2011; Fukumura & Carminati, 2022). Since we are interested in investigating reference to groups of objects, we need to consider more complex contexts with more referents. For this reason, the location descriptions in our study have to be more elaborate than the binary features (e.g., left vs. right, up vs. down) used in Arts et al. (2011), and it is not clear whether the general benefit observed for simple location attributes generalizes to more complex location indexing.

Because of the complexity of the display, the listener also needs to engage in **visual search**, with potentially different underlying mechanisms for reference by common feature compared to reference by location. In visual search, the total set size of objects in the display is known to influence search time for some search targets, but not others. In particular, color is generally considered a pre-attentive feature that does not require the serial deployment of attention to different objects and hence results in similar search times independent of set size. Patterns, on the other hand, could be classified as feature conjunctions, which have been shown to result in search times that linearly increase with set size (Treisman & Gelade, 1980). When multiple (identical) targets have to be found, search time has been shown to increase superlinearly for complex objects (Horowitz & Wolfe, 2001). We thus expect visual search to depend on the type of shared feature, as well as the number of objects to be found.

As the listener does not know how many objects to look for in the case of reference by common feature, termination of search is another relevant concept from the visual search literature. While Treisman and Gelade, 1980 considered an exhaustive search of all objects and reported a linear relationship with set size, later work suggests that terminating a search also depends on object prevalence and internal factors, such as how important it is to find the object (Chun & Wolfe, 1996; Wolfe, 2012). While this work rests on research conducted in single-target trials, on the other end of the spectrum researchers have looked at foraging, where search in one region is terminated mainly for efficiency reasons (Wolfe, 2013). Few authors have looked at search for a smaller number of objects such as the present work. One exception is the use case of radiology or airport security, where the number of targets is typically between 0-2 (Biggs, 2017).

Work on **human-computer interaction** (HCI) has long acknowledged the importance of successful communication between human and machine (e.g. Winograd, 1986). While generating referring expressions (GRE) has received a lot of attention (Dale & Reiter, 1995; Krahmer & Van Deemter, 2012), proposed algorithms concentrate on single object cases (though see Gatt and Van Deemter, 2005) and are mostly influenced by production patterns and evaluated with ratings. More recently, evaluations of GRE have also used human studies using object identification time (Tanaka et al., 2019) and identification success (Kunze et al., 2017) as indices of good referring expressions. The latter two studies are also interesting because they stress the importance of using landmarks to refer to less salient objects. Their findings suggest that referring to object characteristics results in identification failure if these characteristics are not salient enough. For our design, this suggests that reference by location yields better performance when a group is defined by a less salient common feature, such as pattern.

Experiment

This experiment investigates how different reference strategies for the same group of objects – by mentioning a common feature or enumerating coordinates of individual objects – affect response times, identification accuracy and memory encoding. In a within-subjects blocked design, participants received auditory instructions and had to click on all objects in the group. The group size varied from 1 to 4. All 4-object trials included a memory question where participants had to indicate whether a given object was part of the group they have just identified.

We conducted a pilot study with 16 items and 10 participants to validate the task and materials. Based on lower accuracy and increased response times for some items, we excluded purple from the set of critical colors and modified the patterns to achieve better discriminability. We furthermore used the pilot data for sample size calculation by simulation, achieving 90% power for expected effects in clicking accuracy and 99% power for the expected effects in response time. Due to convergence problems on the pilot data caused by data sparse-

ness, we were not able to run a power analysis for memory accuracy. The current study was preregistered in OSF[1].

Participants

We recruited 48 native German speakers (28 male, 20 female, mean age 39.9 years, range 20-70 years) via Prolific. The experiment took approximately 20 minutes to complete. Participants were compensated at a rate of £11.15 per hour.

Design

Our design was a 2 x 2 x 4 design crossing condition (reference by common feature vs. reference by location), feature type (color (blue, green, and yellow) vs. pattern (dotted, striped, and checkered)), and group size (1-4). Condition was manipulated within items and participants, whereas feature type and group size were between-item, but within participants. Condition was blocked. There were four lists that counterbalanced block order (Location block first vs. Referring expression block first) and critical trial conditions (assignment of items to location/referring expression conditions). Participants were randomly assigned to lists. Each block consisted of 43 trials (3 practice trials, 24 critical trials, and 16 fillers), 12 of which (1 practice trial, 6 critical trials, and 5 fillers) were followed by a memory question.

Materials

Visual stimuli. Images consisted of 4×4 grids containing 16 distinct objects that was labeled with alphanumeric coordinates (A1 to D4). The grid size was adjusted to participants' screen size to ensure that it was displayed completely across devices. Objects varied along three feature dimensions: color (blue, yellow, green, purple), shape (heart, cross, crescent, star), and pattern (solid, dotted, striped, checkered). For the 48 critical items, the number of target objects varied from 1 to 4. The target objects were not adjacent. 32 filler items contained 1 to 5 targets that could be adjacent. In the common feature condition, the targets shared the same feature: shape (N = 12), color (purple, N = 2), or pattern (checkered or striped, N = 2). In the location condition, the targets either shared a feature (N = 4), or were random (N = 12). The solid pattern was never a common feature among target objects. We additionally constructed 3 practice trials for each block.

Auditory stimuli. All instructions were recorded by a native German male speaker at a natural speech rate. The referential format of the instruction (spatial coordinates vs. common feature) depended on the experimental block. In the reference by location block, participants heard the coordinates of the cells they should find, e.g.: "Find A4, B3, and D2". The coordinates were always listed in increasing order. In the reference by common feature block, they were asked to find objects with a certain feature, e.g.: "Find the yellow objects", "Find the hearts", "Find the striped objects". All stimuli included around 500 ms of leading and trailing silence. In the referring expression condition, the duration of speech ranged from 1010 to 1753 ms (M = 1394 ms). In the location condition, the duration varied depending on the number of targets: 1 target (M = 1115 ms; range: 1021-1230 ms), 2 targets (M = 1965 ms; range: 1822-2240 ms), 3 targets (M = 2816 ms; range: 2589-3134 ms), 4 targets (M = 3502 ms; range: 3274-3831 ms).

Procedure

On each trial, participants simultaneously saw a grid with 16 objects and heard an instruction specifying the target objects, as shown in Figure 1. They were instructed to click on all target objects as quickly as possible and submit their response. Each trial had a maximum duration of 8 seconds. After 24 trials (28%), if they finished before the timeout, participants were asked a memory question: they saw one object and responded whether it was among the targets from the current trial by clicking "Yes" or "No". On critical trials, the memory object was one of the target objects. On filler and practice trials, they were shown an object that was not on the grid but shared at least one feature with the targets. Before each block, participants completed three practice trials with feedback about their target selection and memory response. After each block, they rated the difficulty of the task (1-5 scale), and at the end of the experiment, they were asked which of the two blocks they found to be more difficult.

The experiment was hosted on LingoTurk, an open-source crowdsourcing server system for psycholinguistic experiments (Pusse et al., 2016).

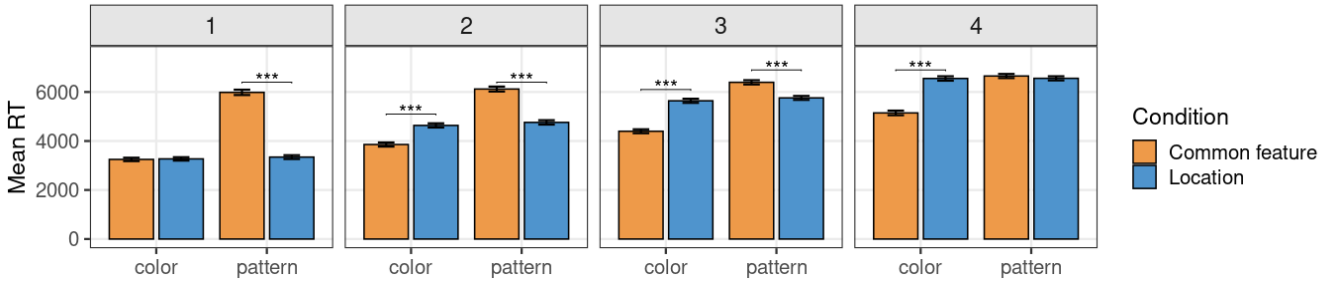
Analysis

All data analyses were conducted using the lme4 package in R (Bates et al., 2015; R Core Team, 2024). Post-hoc comparisons were carried out using the emmeans package (Lenth & Piaskowski, 2025). We fit linear mixed effect models to the log reaction times and logistic mixed effect models to clicking accuracy and memory accuracy. In addition to the effects of interest (condition, group size, and feature type), we included block order and trial number within block as covariates. We used the maximal random effect structure warranted by the data, i.e., random slopes were excluded if they resulted in convergence problems and singular fits. Categorical predictors were effect-coded (condition: Common feature = +1, Location = -1; featuretype: Color = +1, Pattern = -1). Continuous predictors were centered.

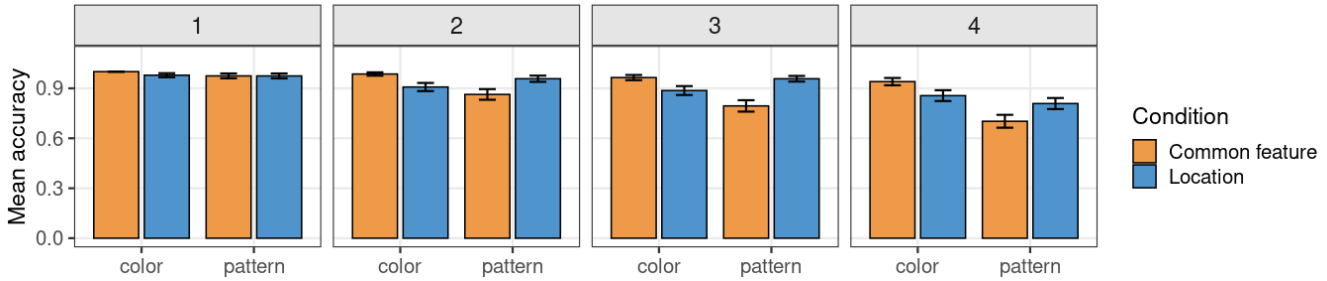
Results

One participant was excluded from analysis because they had experienced technical issues, and 3 items had to be excluded due to errors in the materials. All analyses and graphs below are thus based on 47 participants and 45 items.

[1]The preregistration can be found at: https://osf.io/ux256/overview?view_only=0161f4a6bb9d47f29a15227545bdd0dc. It is currently embargoed to ensure anonymous review but will be made available in the camera-ready version.



(a) Mean reaction times for different group sizes by feature type and condition. Significance annotations indicate Tukey-adjusted pairwise comparisons of estimated marginal means, comparing conditions within each feature type × group size combination.



(b) Mean clicking accuracy for different group sizes by feature type and condition.

Figure 2: Mean reaction time and accuracy in the two conditions for different target group sizes.

Response Time

We excluded all trials where the objects were not identified correctly. The model summary can be found in Table 1. The reaction time analysis revealed significant main effects of group size ($\beta = 0.163$, $t(53) = 23.29$, $p < 0.001$) and feature type ($\beta = -0.111$, $t(58) = -13.77$, $p < 0.001$), indicating that groups that shared a common color were identified faster than those sharing a pattern, and that it took longer to find larger groups. There was a main effect of condition ($\beta = 0.018$, $t(62) = 2$, $p = 0.049$), with search times being on average shorter in the reference by location condition. Condition also interacted with group size ($\beta = -0.064$, $t(45) = -8.86$, $p < 0.001$), with a less pronounced increase in response times for bigger groups in the reference by common feature condition compared to the location condition. The interaction between condition and feature type was also significant ($\beta = -0.105$, $t(39) = -13.84$, $p < 0.001$), with a greater difference between color and pattern in the reference by common feature condition compared to the location condition. Interestingly, there was also a 3-way interaction between condition, number of objects and feature type ($\beta = 0.026$, $t(39) = 3.84$, $p < 0.001$). To understand this interaction, we performed pairwise comparisons using Tukey correction, crossing number of targets and feature type. As shown in Figure 2 (a), if the group is defined by a common color, reference by common feature results in shorter reaction times than enumerating coordinates (all $p < 0.001$), except for group size 1, where the difference is close to zero and not significant (diff = $-0,06$, p=0.101). The benefit of referring by common feature additionally increases numerically with group size, as also apparent in Figure 2a. If, on the other hand, the group is defined by a common pattern, there is an advantage of using coordinates (all $p < 0.001$), except for group size 4, where the difference is again not significant (diff = -0.02, $p = 0.530$).

Clicking Accuracy

Clicking accuracy was only modeled for group sizes larger than 1 to avoid convergence issues due to ceiling effects. In this subset, mean clicking accuracy was 88.42%. A closer inspection revealed that the types of errors differed between conditions: When referring to a common feature, almost all errors were due to missing one or more of the targets. In the location condition, on the other hand, participants often clicked on adjacent objects or locations that sounded similar (e.g., D2 instead of B2), with some instances of target objects being missed. In the inferential analysis, we do not distinguish between the different kinds of mistakes.

The model summary can be found in Table 2. We find a significant main effect of feature type ($\beta = 0.40$, $p = 0.029$), with more errors in the pattern condition, as well as a main effect of group size ($\beta = -0.56$, $p = 0.001$), such that accuracy decreased with increasing group size. The main effect of condition was not significant ($p = 0.653$), nor was the three-way interaction ($p = 0.150$). Interestingly, the interaction between condition and feature type was significant ($\beta = 0.70$, $p < 0.001$). Pairwise Wald tests revealed that accuracy was higher for reference by common feature expressions when the group was defined by a common color (OR = 3.422, $p = 0.007$), whereas accuracy was lower for reference by common feature compared to locations when the group was defined by a common pattern (OR = 0.208, $p < 0.001$).

Table 1: Model of response time (log transformed). The random effect structure for this model was (1 + condition + groupsize_c + featuretype + condition:groupsize_c | participantID) + (1 + condition | itemID)

	Estimate	Std. Error	DF	t value	p value
(Intercept)	8.495	0.022	52.80	393.92	< 0.001
condition	0.018	0.009	62.44	2.01	0.049
groupsize_c	0.163	0.007	53.70	23.29	< 0.001
featuretype	-0.111	0.008	58.37	-13.77	< 0.001
trialnum_c	-0.012	0.003	1740.53	-3.98	< 0.001
first_block	0.005	0.018	44.23	0.26	0.795
condition:groupsize_c	-0.064	0.007	45.01	-8.86	< 0.001
condition:featuretype	-0.105	0.008	39.50	-13.84	< 0.001
groupsize_c:featuretype	0.032	0.006	40.68	5.06	< 0.001
condition:trialnum_c	0.005	0.003	1741.18	1.73	0.085
condition:first_block	0.009	0.006	44.20	1.64	0.109
condition:groupsize_c:featuretype	0.026	0.007	39.61	3.84	< 0.001

Table 2: Model of clicking accuracy. Conditions are abbreviated (condition -> cond, featuretype -> feattype). The random effect structure was (1 + condition + featuretype + groupsize_c | participantID) + (1 + condition | itemID)

	Est	SE	z val	p val
(Intercept)	3.01	0.24	12.40	<0.000
cond	-0.08	0.19	-0.45	0.653
groupsize	-0.56	0.18	-3.18	0.001
feattype	0.40	0.18	2.19	0.029
trialnum	0.09	0.10	0.93	0.351
feattype:cond	0.70	0.15	4.54	<0.000
feattype:groupsize	0.23	0.15	1.49	0.136
cond:groupsize	0.08	0.15	0.53	0.596
feattype:cond:groupsize	-0.20	0.14	-1.44	0.150

Memory Accuracy

The mean accuracy in the memory task for critical items was 70.4%. The model of memory accuracy (Table 3) did not show any significant effects of condition, feature type or their interaction, suggesting that memory encoding of selected items was comparable between conditions, although accuracy is numerically slightly higher in the location condition, as shown in Figure 3.

Table 3: Model of memory accuracy. The random effect structure included was (1 + condition | participantID) + (1 + condition | itemID).

	Est	SE	z val	p val
(Intercept)	0.98	0.17	5.69	<0.001
condition	-0.18	0.14	-1.29	0.195
featuretype	0.01	0.14	0.10	0.922
condition : featuretype	-0.07	0.13	-0.52	0.604

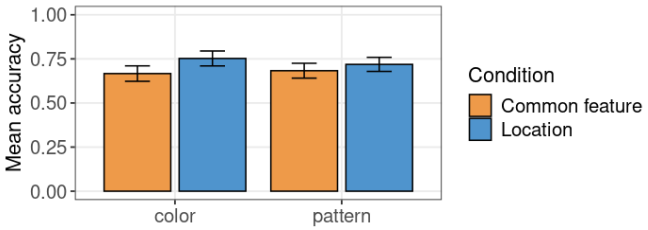


Figure 3: Mean memory accuracy for a selected item in 4-object critical items by feature type and condition.

Subjective difficulty

Finally, there was no difference between the two ways of referring in terms of perceived difficulty. Mean perceived difficulty reported independently after each block was 3.17 (SD = 1.13) and 3.11 (SD = 1.09) out of 5 for referring to a shared feature and enumerating coordinates respectively. When participants were asked to compare both types of references, there was also no clear preference, although a numerically higher number of participants found the referring expression condition more difficult (20 vs. 14; 13 participants indicated that the two conditions were equally difficult).

Discussion

In this study, we investigated two ways of referring to groups of objects: by referring to a shared feature and by enumerating the coordinates of each object. We were interested in whether one type of reference is more effective in terms of response times, identification accuracy, and memory encoding.

The above results give a clear indication that referring to a common feature is not universally better or worse than enumerating individual object coordinates. Instead, we find that the optimal way of referring depends on both the type of common feature and the target group size. In cases where a prominent feature, such as color, defines a group of at least 2 objects, referring to a shared feature leads to faster selections

as well as fewer identification errors. This suggests that an efficient visual search for pre-attentive features is faster than processing coordinates.

However, when the group is defined by a less prominent feature like pattern, we can see a trade-off in response times: for group sizes < 4, there is a clear but decreasing benefit of enumerating individual coordinates, suggesting that a longer utterance can still lead to a faster response when competing against a difficult visual search. With four objects, however, the benefit has almost leveled off. It should be noted that all utterances ended well before the time participants submitted their response, the shortest mean latency being roughly 2 seconds for the one-object case. Therefore, it may be that it is not the actual length of the utterance that causes this leveling-off effect, but rather the increased cognitive load of processing a list of coordinates as opposed to a simpler expression referring to a common feature. Further evidence for this claim comes from the accuracy data, where we observe a marked increase of errors for group size 4 in the location condition. Overall, clicking accuracy for groups that share a pattern is higher in the location condition, and there is no evidence for a similar leveling-off of the difference with increasing group size as found in response times. Finally, we did not find significant differences in recognition accuracy between the two reference types for the group size 4.

The fact that participants took a very long time to find a single object when its pattern was mentioned suggests that participants engaged in an extensive and possibly exhaustive visual search - as the total set size was constant at 16 objects and there were no no-target trials this is in line with previous work (Treisman & Gelade, 1980; Wolfe, 2012). An informal inspection of the delay between selecting the last object and submitting their response confirms this: while this delay was on average around 1 second for all location conditions and the combination of referring expression with color, for pattern we see a much longer delay in the 1-target condition, indicating that all other objects were inspected before deciding to terminate the search. Despite this extensive search, participants were more likely to miss a target with increasing group size, which is in line with previous findings that after the successful detection of one target, additional ones are more likely to be missed (Biggs, 2017).

Our findings have implications for computational cognitive modeling of reference production. The Rational Speech Act (RSA) framework (Frank & Goodman, 2012), which has been used for modeling a variety of pragmatic phenomena, includes a production cost for the speaker but not a comprehension cost for the listener (though see Hawkins et al., 2021 for an example of perspective-taking cost for both interlocutors). Finding that different ways of referring have distinct reaction time and error patterns indicates that, in the context of a complex visual scene, referent identification is associated with a cost for the listener, which suggests that additionally modeling comprehension cost may be warranted. This is also in line with the visual efficiency hypothesis (Rubio-Fernandez, 2021): for the listener, longer utterances are more effective than shorter ones, when they facilitate the search process.

Our results have relevance for human-computer interaction, as they can be informative for developing practical guidelines for efficient communication. Consider, for example, a human controller that oversees a fleet of semi-autonomous transport drones (see e.g., Knieriemen et al., 2025). In the case of a critical situation that requires the swift identification of a subset of drones, our results suggest that a good strategy would be to highlight relevant drones with a salient color on the display and refer to them verbally as, e.g., "the highlighted drones". In a less time-sensitive situation, where highlighting should be avoided in order not to visually overload the interface, an enumeration of coordinates will lead to faster and more accurate identification of the relevant targets. An example of such a situation could be that the system identified a number of drones for which the controller should check the battery status to update their situation model.

Limitations and Future Work

While this study is the first to systematically compare identification time and accuracy for groups of objects with different instructions, this experiment has some limitations. First of all, the use of basic shapes and simple features like color and pattern may limit the generalizability of our findings to more complex real-world scenarios. Secondly, as we conducted our experiments online, our primary measure is response time, which conflates identification time with search termination, making it difficult to compare our findings directly to previous work. Future work could conduct an eye-tracking study, which would provide more fine-grained information about which objects are inspected at all and when visual search is terminated. Finally, our finding regarding groups that share a less salient feature remains somewhat inconclusive: we see that the benefit of enumerating coordinates decreases with group size, but our largest groups (four objects) do not show a clear benefit of using a referring expression. This benefit may emerge with larger group sizes.

Acknowledgments

We thank the anonymous reviewers for their thoughtful comments and suggestions. This work was partially funded by DFG grant 389792660 as part of TRR 248 – CPEC, see https://perspicuous-computing.science.

References

Arts, A., Maes, A., Noordman, L., & Jansen, C. (2011). Overspecification facilitates object identification. *Journal of Pragmatics*, *43*(1), 361–374. https://doi.org/10.1016/j.pragma.2010.07.013

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Biggs, A. T. (2017). Getting satisfied with "satisfaction of search": How to measure errors during multiple-target visual search. *Attention, Perception, & Psychophysics*, *79*(5), 1352–1365.

Chun, M. M., & Wolfe, J. M. (1996). Just say no: How are visual searches terminated when there is no target present? *Cognitive psychology*, *30*(1), 39–78.

Dale, R., & Reiter, E. (1995). Computational Interpretations of the Gricean Maxims in the Generation of Referring Expressions. *Cognitive Science*, *19*(2), 233–263. https://doi.org/10.1207/s15516709cog1902_3

Elsner, M., Clarke, A., & Rohde, H. (2018). Visual complexity and its effects on referring expression generation. *Cognitive science*, *42*, 940–973.

Engelhardt, P. E., Barış Demiral, Ş., & Ferreira, F. (2011). Over-specified referring expressions impair comprehension: An ERP study. *Brain and Cognition*, *77*(2), 304–314. https://doi.org/10.1016/j.bandc.2011.07.004

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998–998.

Fukumura, K., & Carminati, M. N. (2022). Overspecification and incremental referential processing: An eye-tracking study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *48*(5), 680–701. https://doi.org/10.1037/xlm0001015

Gatt, A., & Van Deemter, K. (2005). *Towards a psycholinguistically-motivated algorithm for referring to sets: The role of semantic similarity* (tech. rep.). Technical report, TUNA Project, University of Aberdeen.

Grice, H. P. (1975). Logic and conversation. In *Speech acts* (pp. 41–58). Brill.

Hawkins, R. D., Gweon, H., & Goodman, N. D. (2021). The division of labor in communication: Speakers help listeners account for asymmetries in visual perspective. *Cognitive science*, *45*(3), e12926.

Horowitz, T. S., & Wolfe, J. M. (2001). Search for multiple targets: Remember the targets, forget the search. *Perception & Psychophysics*, *63*(2), 272–285.

Knieriemen, N., Hirsch, A., Moiz Sakha, M., Daiber, F., Kolb, H., Hüning, S., Wiehr, F., & Krüger, A. (2025). Seeing, hearing, feeling: Designing multimodal alerts for critical drone scenarios. In *Proceedings of the 27th international conference on multimodal interaction* (pp. 456–465). Association for Computing Machinery. https://doi.org/10.1145/3716553.3750784

Koolen, R., Gatt, A., Goudbeek, M., & Krahmer, E. (2011). Factors causing overspecification in definite descriptions. *Journal of Pragmatics*, *43*(13), 3231–3250.

Krahmer, E., & Van Deemter, K. (2012). Computational generation of referring expressions: A survey. *Computational Linguistics*, *38*(1), 173–218.

Kunze, L., Williams, T., Hawes, N., & Scheutz, M. (2017). Spatial referring expression generation for hri: Algorithms and evaluation framework. *AAAI Fall Symposia*, 27–35.

Lenth, R. V., & Piaskowski, J. (2025). *Emmeans: Estimated marginal means, aka least-squares means* [R package version 2.0.1]. https://rvlenth.github.io/emmeans/

Pechmann, T. (1989). Incremental speech production and referential overspecification.

Pusse, F., Sayeed, A., & Demberg, V. (2016). Lingoturk: Managing crowdsourced tasks for psycholinguistics. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, 57–61.

R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Rubio-Fernandez, P. (2021). Color discriminability makes over-specification efficient: Theoretical analysis and empirical evidence. *Humanities and Social Sciences Communications*, *8*(1), 147.

Tanaka, M., Itamochi, T., Narioka, K., Sato, I., Ushiku, Y., & Harada, T. (2019). Generating easy-to-understand referring expressions for target identifications. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5794–5803.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive psychology*, *12*(1), 97–136.

Winograd, T. (1986). A language/action perspective on the design of cooperative work. *Proceedings of the 1986 ACM conference on Computer-supported cooperative work*, 203–220.

Wolfe, J. M. (2012). When do i quit? the search termination problem in visual search. *The influence of attention, learning, and motivation on visual search*, 183–208.

Wolfe, J. M. (2013). When is it time to move to the next raspberry bush? foraging rules in human visual search. *Journal of vision*, *13*(3), 10–10.