Modification strategies for discriminating among referents in the presence of distractors: An analysis of large-scale production data

Marina Bolea, Louise McNally, and Peter R. Sutton

Universitat Pompeu Fabra
{marinabcent@gmail.com, louise.mcnally@upf.edu, peter.r.sutton@icloud.com

Abstract. Controlled experimental studies have repeatedly established that speakers adjust the referential expressions they use to take into account other candidate referents in a given context. These studies have been used to probe the extent to which reference strategies are communicatively "efficient" and indeed whether the efficiency of an expression is best defined in terms of the minimal number of entailments needed to differentiate a target referent or in some other way, e.g., in terms of the speed with which an expression helps interlocutors coordinate on the target. Most of these studies have used decontextualized images which have, moreover, focused on very specific sets of potentially discriminating features, such as color or size. We present an analysis of data collected in a task that involved distinguishing target from distractor referents in naturalistic images; our goal was both to test the ecological validity of previous experimental studies as well as to identify additional discriminating features that had not been explored in previous research. Our results provide new evidence for the salience of entity *parts* and reveal finer-grained information about the relative uses of and dependencies between visual and other features, which can inform future experimental studies of how speakers communicate their targets of reference.

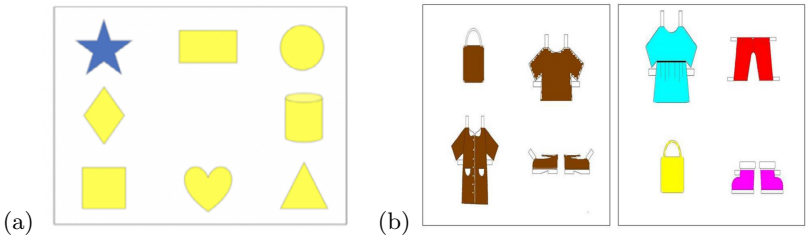
Keywords: referring expressions · modification strategies · 'redundant' modification · colour modification · reference to part structure

1 Introduction

It is a central claim of theories of linguistic communication inspired by Grice's Cooperative Principle that, in order to be efficient, speakers should balance informativity and brevity ([9], [11]). For example, in order to uniquely identify the object in the upper left hand corner of image (a) in Figure 1, it should sufficient to utter "(the) star", without any mention of its colour, as there are no other stars in the context; the same applies to all of the entities in the blocks in images (a) and (b) in Figure 1 (on which, more below).

For over a decade (see [7], [19], [3], for early examples), controlled experimental studies have repeatedly established that speakers adjust the referential expressions they use to take into account other candidate referents in a given

Fig. 1: Stimuli from (a) [20] and (b) [21].



context. Broadly, the results show that speakers consider factors such as the relative cost and informativity of candidate messages, as well as inferences about their interlocutors' knowledge state, and tend to behave in ways that conform with the Gricean vision of efficient communication.

However, alongside these results, there is also a long line of studies that reveal that speakers are *not* always maximally efficient, if efficiency is defined strictly in terms of minimizing unnecessary entailments (see [4] for a review). For example, [18], [23], [6], [20], [21], [13], and [26], among others, have shown that participants will use colour adjectives redundantly, such as "blue star" to refer to the star in Figure 1, image (a). This is particularly likely to happen when the entity is of an atypical colour ([20]; see [15] for similar results involving shape and material); or the more the colour is contrastive and pops-out, and so especially salient, as in image (a) or the right-hand block in image (b) in Figure 1, as opposed to the left-hand block in the same image ([20], [21]). Other studies point to a similar salience effect in the use of redundant spatial expressions – for example, "the ball on top of the cube" to indicate the ball in image (a) in Figure 2 ([25]). Moreover, features of the visual scene, such as its overall complexity or the number of distractors, have also been shown to influence redundant colour modifier use (e.g. [2], [13], [5]).

A related general observation involves the specificity of the noun chosen for reference. [12] and [8] have shown that speakers will use overinformative nouns, for example, a subordinate-level term such as "Dalmatian" when "dog" would be sufficient, especially if the referent is more typical for the subordinate-level than for the basic-level category; [8] also documented the use of basic-level terms such as "dog" when a less granular, superordinate-level term such as "animal" would be sufficient (see, e.g., image (b) in Figure 2).

These latter studies have led to efforts to redefine the notion of efficiency or explain under what circumstances "inefficient" information is communicatively useful (e.g., [21], [4]). For example, [21] concluded that the visual salience of a contrasting colour can help a hearer more rapidly identify the referent in a heterogeneous visual scene; [25] see a similar role for spatial relations, especially topological ones, in locating objects in more complex visual scenes. Thus, efficiency should be defined not only in terms of the information *qua* entailment

needed to identify an object, but also in terms of the time and effort needed to identify it.

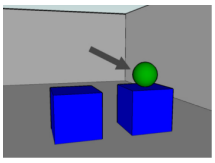
(a)  (b) 

Fig. 2: Examples of the stimuli used in (a) [25] and (b) [8].

These results are intriguing, but given that they involve controlled experimental designs, they have limited ecological validity. The images are often drawn or decontextualized; images with multiple entities in them are relatively simple; and the contrasting features that have been studied have focused largely on colour, size, shape, and spatial configuration. The goal of our study is to contribute to this body of literature by providing an analysis of a more naturalistic set of data, drawn from [17]. In addition to allowing us to check the ecological validity of previous studies which have examined efficiency in referring expression choice, naturalistic images both provide visual context, which can influence that choice and, more relevantly for our purposes, afford speakers a wider set of distinguishing features to choose from than is typically found in highly controlled experimental tasks. As we shall discuss, one finding we make is that, especially when target and distractor images are similar, speakers often refer to parts of objects or nearby objects to distinguish targets from distractors (see, e.g., [1]; see also [10], especially Chapter 3, for discussion of the role of visual background in influencing referring expression choice.)

Specifically, we present the results of a quantitative and qualitative analysis of modifier use collected by, but not analyzed in, Mädebach et al. ([17]), a study of referring expression production in a referent-identification task involving naturalistic visual images such as that in Fig. 3 (see section 2.1 for details). These images allowed us to look at preferences speakers manifest *when they have a choice*. We focused specifically on uses of modifiers, setting aside an analysis of variation in lexical choice, as the latter was the goal of Mädebach et al.'s study (again, see below).

We were interested in three main questions:

1. **Which features do speakers use** in naturalistic settings to distinguish referents from distractors? In particular, is there evidence for the use of some kinds of features that have not been the primary focus of previous experimental studies?
2. **What dependencies, if any, are there between the features expressed**, if more than one feature is chosen?

3. **To what extent do we find evidence of 'redundant' modifier usage** when speakers have free choices over referring expressions in naturalistic settings (and do the results cohere with those in the experimental literature)?

In addition, in all cases we also checked the distribution of the modifiers before vs. after the noun.

Our goal at this point was not to put forward a theory for why speakers make specific choices with respect to referring expressions but rather constituted what we consider to be a necessarily preliminary step. Especially in addressing questions 1 and 2, we sought first to clarify the empirical landscape when it comes to what kinds of modifiers speakers use when faced with more open ended choices of expressions in naturalistic settings.

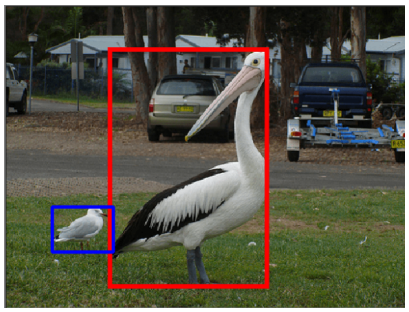


Fig. 3: An example of an image from [17], with the target entity bounded by a red box and the distractor by a blue box.

With respect to question 3, our dataset allowed for redundant uses of modifiers to appear in at least two ways: (i) when a modifier accompanies a noun which would be sufficient on its own to distinguish the referent from the distractor (e.g., using "the big pelican" when presented with Figure 3, where the distractor is a different kind of bird); (ii) where e.g., colour plus another modifier is used when the latter alone would be sufficient.

Our results can be summarized as follows. First, in addition to using visual features (e.g. colour or size) and position or orientation, speakers also relied on less-studied features such as descriptions of the event, process or state in which the referent participates and, especially, reference to salient parts of the referent or other objects in the vicinity. These findings can inform the design of future experimental studies with more controlled images, for instance, to determine under what conditions features like event descriptions or parts of objects may be selected by speakers over, for instance, descriptions of colour or shape.

Second, we uncovered some dependencies between these features. For example, if a speaker introduces visual information[1] about the target object such as

[1] As we clarify in section 2.2, we use 'visual information' in a technical sense that corresponds approximately to basic physical features and excludes, e.g., describing

its predominant colour, or overall size or shape (e.g., *big/white bird*), such that this information differentiates it from the distractor, this tends to preclude the use of other information. However, if either a part of the target object or another related or nearby object is referred to, this is far more likely to be accompanied by some ancillary information about that part or other object. Since parts of objects or nearby objects are less visually prominent than the main objects in the main target or distractor images, this provides some evidence that some notion of salience or prominence may be an important element in understanding speakers' uses of referring expressions.

Finally, we also found some limited evidence of redundant uses of colour adjectives in participants' descriptions of objects in naturalistic settings. The sample found in our dataset is small, and so we cannot draw any strong conclusions from it, but the rates at which speakers use redundant colour adjectives is consistent with those reported in [22]. Also, we found one type of redundant use of color adjectives that has not, to our knowledge, been attested in previous experimental studies but rather reflects the complexities of photo-realistic images: For an image with two black shirts, where the target had no logo, a number of speakers used "black shirt without a logo" even though neither *black* nor *shirt* distinguished the target image from the distractor.

2 Method

2.1 The Mädebach et al. dataset

Mädebach et al. [17] built their study on Silberer et al.'s [24] ManyNames dataset of referring expression choices in naturalistic visual contexts. Via Amazon Mechanical Turk (AMT), [24] had participants label entities in images drawn from Visual Genome ([14]); these entities were marked in red bounding boxes, and the labeling instructions did not involve distinguishing the entity from any potential distractor. On average, the ManyNames dataset provides 31 names per image.

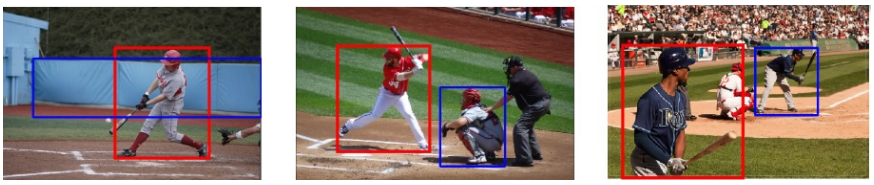
[17] selected 72 images from ManyNames in which the target (again, in a red bounding box) was accompanied by a distractor entity (in a blue bounding box); 97 participants (also via AMT) were asked to label the targets. The images were equally divided among three conditions: *no-competitor*, *lexicon-sufficient*, and *syntax-necessary* (see Fig. 4). Here we limit discussion to the latter two conditions, which were the only ones we analyzed.[2]

---

the position or orientation of an object relative to the whole image or some other object, a description of eventualities in which objects are participants. It also excludes descriptions of specific parts of objects.

[2] According to [17], the determination of whether an image fell into the lexicon-sufficient or syntax-necessary condition was made manually by the authors. A reviewer correctly points out that this method does not guarantee that all participants were familiar with the lexical items that would distinguish target from distractor in the lexicon-sufficient condition. That said, all of the images used in the study were ones for which, in the original ManyNames dataset, both more general names (e.g.

In the *lexicon-sufficient* condition, there was a lexical item that could distinguish the target from the distractor (e.g., *batter* in Fig. 4 (b)); in the syntax-necessary condition, there was no such lexical item, and therefore some sort of syntactic modification was needed to distinguish the two entities (e.g. *the batter on the grass* in Fig. 4 (c)). Participants were told to type into an input box an expression that would allow an interlocutor to identify the entity in the red bounding box; they were given *the* as a prompt. Participants were instructed not to use the bounding box and to avoid using spatial expressions such as *on the left*, which presuppose a visual perspective that might be distinct for the interpreter. Although it was predicted that speakers would opt for a single lexical item in the *lexicon-sufficient* condition (and indeed, they tended to do so), in 38.6% of the trials in this condition participants used either syntactic modification or a combination of a distinct lexical item and syntactic modification to identify the target. These items (825 in total) as well as the expressions produced in the syntax necessary condition (1439 in total), form the basis for our study.



(a) no-competitor (b) lexicon sufficient (c) syntax-necessary

Fig. 4: Examples of the three conditions. Images and categories as assigned by Mädebach et al. ([17]: Figure 1)

2.2 Annotation of the data

The data from Mädebach et al. was part-of-speech tagged and dependency parsed as well as corrected for typos. We further annotated the data using a combination of automated and manual methods. First, for each referring expression, we used a Python script to sort prenominal modifiers (everything that appeared between *the* and the head noun, e.g. *large yellow* in *the large yellow vehicle*) from postnominal modifiers (everything after the head noun, e.g. *with the black wheel* in *the cab with the black wheel*). Our search strategy was designed to prioritize identifying as many potentially relevant data points as possible, even if this meant initially making incorrect identifications (e.g., in *the cab with the black wheel*, *black* was identified by the script as a colour modifier of *cab*). We adopted this strategy so as not to lose responses that were fragmentary, or written in a

player) and more specific names (e.g. *batter*) were produced by at least some participants. The full set of test items from Mädebach et al.'s study is available on OSF: https://osf.io/p3jt5/.

slightly compressed or not entirely grammatical form, e.g. *the chair with neat blanket*, *the black shirt no logo*.[3] In English, we hypothesised that such a contrast could show up between attributive and relative claause uses of modifiers (*the blue chair* vs. the chair that is blue). Fortuitously, as will be seen below, this strategy proved to reveal the fact that sometimes colour or other modifiers of a part of a target proved to be salient for discriminating that target from a competitor. All incorrect examples retrieved by this method were subsequently manually corrected.

Second, we created a set of attribute lists. These were initially manually seeded with common attributes (for example, we populated the colour attribute with a list of common colours) and fairly fine-grained. The full lists of values for each attribute are available at `github.com/peter-sutton/referring_expressions`. The attribute categories included `colour`, `shape`, `size`, `age`, `orientation`, and `position`; the full set with illustrative examples appears in Appendix A. We then used a Python script to search the pre- and postnominal modifier fields for the words in our attribute lists and, whenever one of those words was found, add the corresponding attribute label to the entry for the referring expression. Some examples of data points and the attributes assigned to them are given in Table 1.

Data point	Prenominal Attributes	Postnominal Attributes
the player swinging$_1$ the racket$_2$	∅	action$_1$ part-or-other-object$_2$
the chair in front$_1$ of green$_2$ lamp$_3$	∅	position$_1$ colour$_2$ part-or-other-object$_3$
the larger$_1$ brown$_2$ chair	size$_1$ colour$_2$	∅
the Lobster Pot$_1$ building	text$_1$	∅
the small$_1$ cow with a very white$_2$ face$_3$	size$_1$	colour$_2$, part-or-other-object$_3$

Table 1: Examples of data points tagged with attributes either prenominally or postnominally. The subscripts associate each modifier with the corresponding attribute.

Once this initial pass was made, we manually reviewed the referring expressions for which there were no tags. As a result of this process, in some cases we added additional expressions to the attribute lists (e.g. *dark* in *the dark desk*

[3] Another motivaation for sorting modifiers into pre- and postnominal was that linear order relative to the noun has been found to be relevant to the rate at which speakers use redundant modifiers especially color terms when contrasting English and Spanish [20].

as `colour`). We also created additional categories: `text` (for when participants used any text in the image as part of their description, e.g. *Lobster Pot* in the penultimate row of Table 1, which figured on a sign on the target object); `visual pattern` (e.g., *decorative* in *the decorative shower curtain*); `other adjective` (e.g., *full* in *the man with a full beard*), `with` (for any responses with a *with PP*), `part_or_other_object` (where a part of the target object or else another object in its vicinity was named, e.g., *the person with the racket lower than his knee*); `action` (*the flying bird*); and `state` (*the buttoned up black shirt*).

After the initial pass, we then manually checked the tagging results, and where necessary, adjusted the lists of terms associated with each attribute. We then re-ran the automatic tagging and manually checked the data once again.

Our methods for deriving initial attribute lists depended on the category. For instance, for `colour`, we used freely available extensive online lists of colour terms. However, for `part_or_other_object`, we took every piece of data with a *with*-PP and extracted a list of nouns in the PP. After first pass tagging, we found that this method both over and under-generated the `part_or_other_object` tag. For instance, some colour terms were included (e.g., *silver*), given that they were parsed as nouns, and some nouns that often named other objects in some scenes were missed, since they were not used in a *with*-PP (e.g., *computer* in *the computer desk* used to refer to a desk with a computer on it).

We also found some expressions which, contrary to expectation, did not contain any modifiers. For instance, some participants responded with *the airplane* in a scene where both the target and the distractor were passenger airliners. These were excluded from the final analysis.

Finally, we then grouped the attributes into four coarse labels and tagged each expression with the relevant labels for the type of information used to modify the head noun: 1) `visual` information included modifiers expressing colour, material, shape, size, age, visual patterns, and text; 2) information regarding the `position or orientation` of the target (e.g., *the horizontal plane*, *the chair against the wall*); 3) descriptions of actions or states (`eventuality`, e.g., *the man doing a trick*, *the brightly lit restaurant*) and 4) mentions of `parts` of the target entity or of `other objects` within the bounding box (e.g., *the chair with armrests*, *the table with flowers*). See Appendix A for examples.[4]

These four tags (visual information, position or orientation, eventuality, and part (or other objects) are not mutually exclusive, as one expression can include more than one type of information (e.g., *the black train with red wheels* was tagged as containing both `visual` and `part` information). For each expression,

[4] As noted by a reviewer, all of the above categories could be classified as 'visual information' insofar as the data were collected in a visual task. While we do not deny this, here we use 'visual information' in a slightly rarefied sense, for those visual features of a whole object that are non-specific to the image as a whole, the perspective from which the image was taken or other objects in the image (thus, distinct from position and orientation information), not related to a description of eventualities, and not specifically referring to some particular part of the target object.

we also noted whether the modifier appeared prenominally (Left of the head noun) or postnominally (Right of the head noun).

3 Data analysis

In this section we include the results of three analyses: first, we present the frequencies and distribution of the four attribute tags in the *lexicon-sufficient* and *syntax-necessary* conditions. Second, we reveal the dependencies between tags through a Bayesian analysis of the data. Finally, we analyse the use of redundant colour adjectives in a subset of our dataset.

3.1 Tag frequency and distribution

The first of our main questions concerned which features speakers use in naturalistic settings to distinguish referents from distractors and whether there was evidence for the use of features not previously explored in experimental studies. To assess this we looked at the frequencies of tags across the two conditions both pre- and postnominally. The results are summarized in Table 2. Recall that the total number of referring expressions analysed for the *lexicon-sufficient* condition was 825, and for the *syntax-necessary* condition, 1439. Fig. 5 summarizes the raw counts of the different coarse categories of modifiers in pre- and postnominal position, in both the *lexicon-sufficient* and *syntax-necessary* conditions.

	lexicon-sufficient		*syntax-necessary*	
	Left	Right	Left	Right
Visual	517	84	619	255
Position/Orientation	20	86	177	197
Eventuality	32	76	108	142
Part/Other Object	1	147	109	385
Total	575	389	1013	979

Table 2: Frequencies of attribute tags for the *lexicon-sufficient* and *syntax-necessary* conditions.

Looking across both left and right contexts, in the ***syntax-necessary*** condition, we find that modifiers providing `visual` attributes predominated. Indicating a `part` of the target or another object within the bounding box was the second most frequent type of modification. If we distinguish Left vs. Right contexts, we see that, prenominally, not only is `visual` information the most frequent tag, but there is also a wide gap between `visual` and all other types of information. Postnominally, `part/other object` is the most common tag, followed by `visual`, `position/orientation` and `eventuality` tags.

In the ***lexicon-sufficient*** condition, `visual` information was also the most frequent attribute when syntactic modification was used, with the gap between

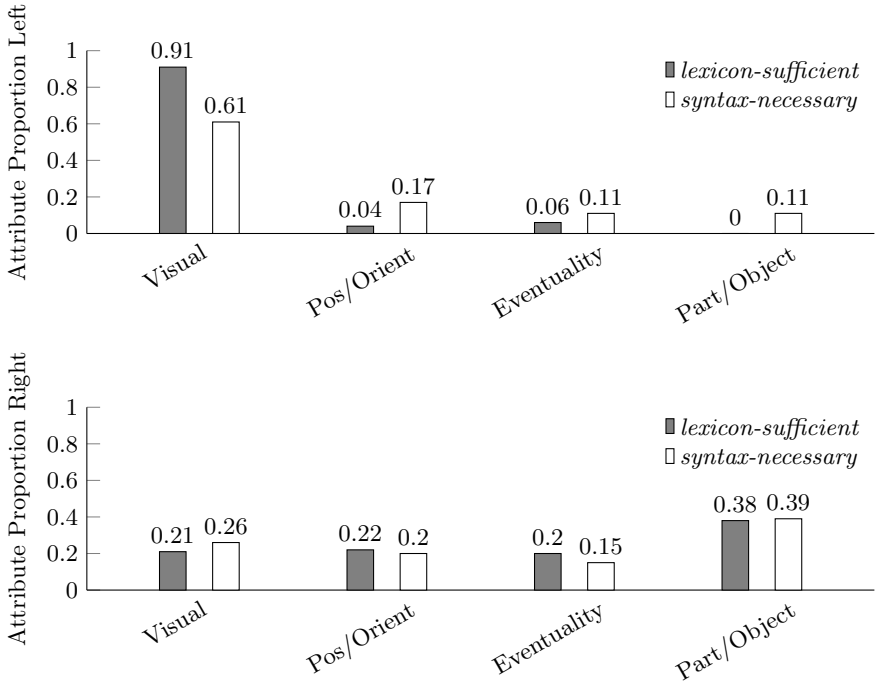


Fig. 5: Proportions of attributes used by position in the NP for the *lexicon-sufficient* and *syntax-necessary* conditions, for visual, position/orientation, eventuality, and part or other object tags.

this tag and all others being wider than in the *syntax-necessary* condition. `Visual` features dominated in prenominal modification, with all other tags being used very rarely. Postnominally, `part/other object` was the most frequent attribute, followed by `position/orientation`, `visual` and `eventuality` information, with the differences between feature frequency being weaker than in the *syntax-necessary* condition.

Discussion Our results echo previous findings, e.g. regarding the salience of colour and other visual features, but we bring additional new observations that should be investigated in future research, especially the usefulness of referring to parts, an underrepresented observation in previous studies (though see [1,16]). In general, our findings point to the relevance of salience when referring to entities in a naturalistic visual context, and the need to move beyond simple notions of informativity and efficiency.

3.2 Quantitative analysis of dependencies between tags

In order to establish and quantify any dependencies between modification strategies (our second main question), we carried out a Bayesian analysis. We computed the conditional probabilities of finding left and right tags if participants used a given left or right tag. For example, we computed the conditional probability of finding `visual` information on the left if participants used `part` information on the right: Pr(Visual-Left|Part-Right).

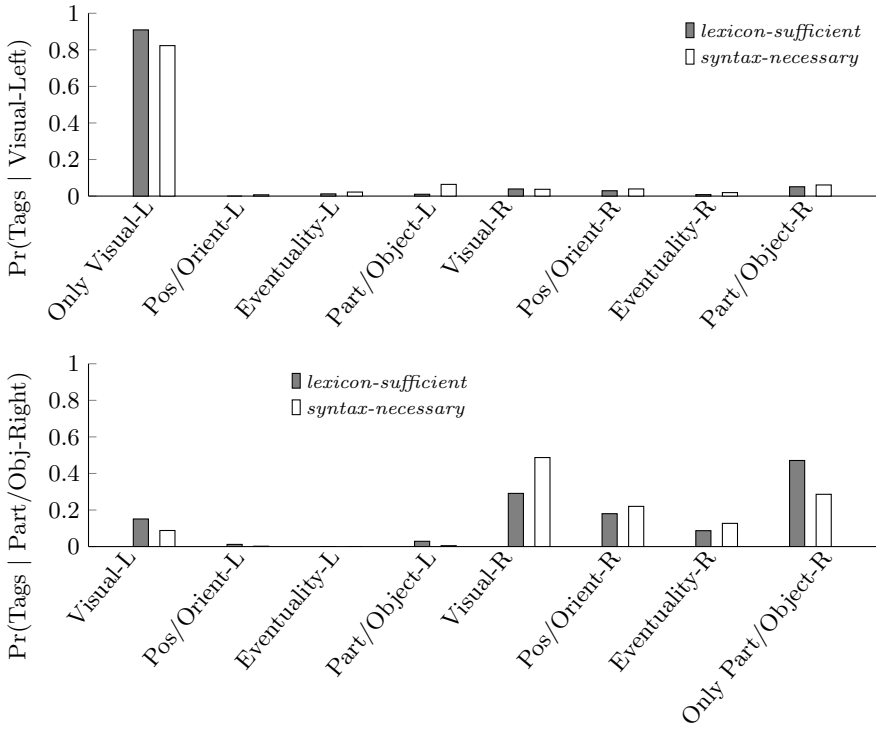


Fig. 6: Conditional probabilities for $Pr(Tag \mid Visual\ Left)$ (top) and $Pr(Tag \mid Part/Other\ object\ Right)$ (bottom) for the *lexicon-sufficient* and *syntax-necessary* conditions. NB: Aside from Only Visual/Only Part, the categories conditioned upon Visual-Left and Part/Object-Right are not disjoint, since tags can co-occur.

Results The analysis reveals that left and right modification tend to exclude one another – few people put information on both the left and the right. Additionally, `visual` information on the left tends to exclude any other information on the left in both conditions: It occurred with additional information in only $\approx 9\%$ of cases in the *lexicon-sufficient* condition and $\approx 12\%$ of cases in the *syntax-necessary*

condition (top in Fig. 6). In contrast, `part/other object` information on the right correlates with the use of other information on the right (bottom in Fig. 6). The effect is stronger in the *syntax-necessary* condition ($\approx$ 73% of cases) than in the *lexicon-sufficient* condition ($\approx$ 53% of cases).

Discussion These results expand what we found in section 3.1 with regard to the salience of `visual` features: they are not only the most frequent attribute, but they almost always appear on their own. We also note a tendency for `parts` to be supported by (or to support) other features. This happens less often in the lexicon-sufficient condition; we speculate that this might be because, in this condition, there are more features to distinguish target and distractor, making reference to a specific part of the target entity less helpful, unless the part is very salient on its own, without need for other features as support.

3.3 'Redundant' uses of colour adjectives

Finally, we also examined the dataset to establish if it provides any evidence for 'redundant' referring expressions similar to those attested in previous studies with colour terms. Even though the images chosen were not designed to test for this possibility, we wished to explore whether such uses of colour terms in particular also emerged in a naturalistic image setting. To do this, we restricted ourselves to the *syntax-necessary* condition, since it yielded more syntactically complex responses from participants.

Results The strongest correlation between our attribute tags were a combination of `part/other object` and `visual` information (which includes colour terms) postnominally (see Fig. 6, bottom). For instance, *the girl with large earrings* was tagged with `part/other object` in virtue of the mention of earrings, and as `visual` due to a specification of their size.

In the *syntax-necessary* condition, 199 data points were tagged with both `part/other object` and `visual` information to the Right of the noun. Approximately half of these were associated with four images, given in Fig. 7.

Of these four images, only the ship and polo shirt images plausibly involved a redundant use of colour adjectives. For the frisbee player and skateboarder images, shirt/t-shirt and shoe colour clearly was not redundant in distinguishing the target from the distractor image.

For the ship image, 24/52 participants referred to the life preserver that was visible only on one of the ships. Of these, 17 (71%) mentioned its colour.

For the polo shirt image, of the 67 responses, approximately one quarter (16) of these mentioned the lack of a logo on the target shirt. Of these, half (8) also mentioned that the shirt was black (e.g., *the black shirt no logo*). This also counts as the use of a redundant colour term, since both the target and distractor polo shirts were black.

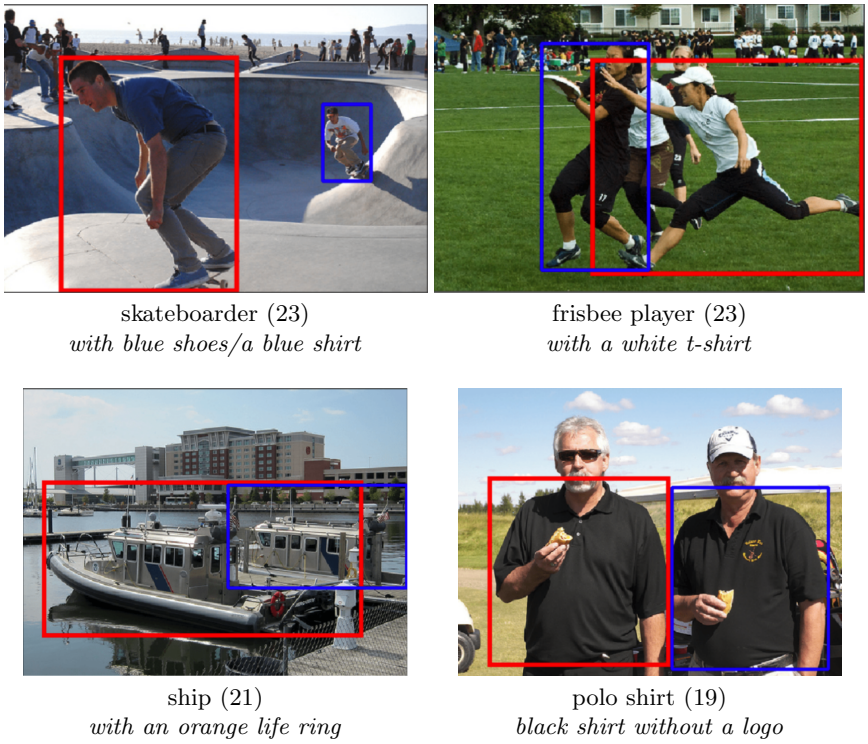


Fig. 7: Image, number of datapoints and typical postnominal information for the four most frequent images with both part/other object and visual information to the right of the noun. Numbers in parentheses indicate the number of datapoints for each image with both tags. The italic text indicates what was typically referred to by participants.

Discussion Our evidence for 'redundant' uses of colour terms was limited (based on only two images and on 40 participant responses). However, insofar as we can draw conclusions from this limited evidence, it is notable that the entity on the ship in the target box that was referred to, the life preserver, stood out in the image, given that it was the only brightly coloured entity. This lends some credence to the view in [20], that what might be considered 'inefficient' uses of colour terms actually serve the purpose of quickly and efficiently drawing the gaze of an interlocutor to the right part of a visual scene, especially if the scene is complex. In our dataset 71% of participants who mentioned the life preserver also mentioned its colour. This is comparable to the rates reported in lab settings in [22], when the entity has a unique colour in the scene, as with the orange of the life ring.

That said, it has been found that, when a colour is predictable based on the type of object (e.g., *yellow* for *banana*), the use of colour terms to refer to this object are lower than when the colour is not predictable [23]. Given that bright

orange/red are predictable colours of life preserver rings on boats, this could suggest an artificially low use of the colour term in this case.

For the polo shirt responses, the use of *black* is somewhat surprising. Unlike the typical cases from the literature where the is only one N (e.g., one star as in Fig. 1), where there was a 'redundant' use of *blue*, for this image, *black* was 'redundant' in a different way: both shirts were black, and so *black* did not even distinguish between the shirts. Perhaps this is why only half of those that referred to the lack of logo on the shirt also mentioned its colour. As for why anyone would mention the colour of the shirt, we see two possible explanations. The first is similar to the life preserver case: The blackness of the shirt does quickly and efficiently draw one's gaze to the two shirts (and additionally perhaps to the sunglasses of the man on the left). So, even if *black* does not distinguish between the two shirts, the mention of the colour does reduce one's search down to two or three entities in the scene. The second possible explanation for why the colour term was used in this case, could, we surmise, also be a result of the domain of reference being *clothing*. As observed in [20], descriptions of clothes include colour terms at higher rates than in other domains.

3.4 Conclusions and future work

We have presented an analysis of speakers' preferences for different types of attributes in a discrimination task. Our results point to visual information (as we have circumscribed this category) being the preferred type of attribute, as well as to it being considered sufficient by speakers, as it rarely co-occurs with other types of information. Our study also suggests that the use of parts of an object to discriminate between that object and distractors is more prominent than previous work suggested. This use of parts, however, is often accompanied by the use of other attributes, signaling that parts are possibly being used as support for other features. We have also found evidence of colour adjectives being used redundantly, in line with previous work.

Future research should consider the development of visual stimuli that are both naturalistic and controlled. These types of stimuli would be particularly useful for gaining a better understanding of the role of parts in a discrimination task. While our study clearly points to mentioning parts as a prominent strategy, it is unclear to what extent this is a result of parts being salient attributes, as opposed to being a last resort, used only when there is no other type of information available. A more controlled dataset would make it possible to examine this issue.

Finally, future work should also consider potential crosslinguistic variation. We might find significant differences in the frequencies of the types of attributes, as well as in their dependencies, if we look at languages where adjectives are typically postnominal (cf. [20]).

Acknowledgements

References

1. Clarke, A.D., Elsner, M., Rohde, H.: Where's Wally: the influence of visual salience on referring expression generation. Frontiers in Psychology **4** (2013). https://doi.org/10.3389/fpsyg.2013.00329
2. Davies, C., Katsos, N.: Are speakers and listeners only moderately Gricean? an empirical response to Engelhardt et al. (2006). Journal of Pragmatics **49**(1), 78–106 (2013). https://doi.org/10.1016/j.pragma.2013.01.004
3. Degen, J., Franke, M., Jäger, G.: Cost-based pragmatic inference about referential expressions. In: Proceedings of the Thirty-Fifth Annual Conference of the Cognitive Science Society. pp. 376–381 (2013)
4. Degen, J., Hawkins, R.X.D., Graf, C., Kreiss, E., Goodman, N.D.: When redundancy is useful: A Bayesian approach to "overinformative" referring expressions. Psychological Review **127**(4), 591–621 (2020). https://doi.org//10.1037/rev0000186
5. Gatt, A., Krahmer, E., van Deemter, K., van Gompel, R.P.G.: Reference production as search: The impact of domain size on the production of distinguishing descriptions. Cognitive Science **41**(Suppl 6), 1457–1492 (2017). https://doi.org/10.1111/cogs.12375
6. Gatt, A., van Gompel, R., Krahmer, E., van Deemter, K.: Non-deterministic attribute selection in reference production. In: Workshop on Production of Referring Expressions: Bridging the gap between empirical, computational and psycholinguistic approaches to reference (PRE-CogSci'11). pp. 1–7 (2011), `https://albertgatt.github.io/dl/precogsci2011.pdf0`
7. Golland, D., Liang, P., Klein, D.: A game-theoretic approach to generating spatial descriptions. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. pp. 410–419. Association for Computational Linguistics, Cambridge, MA (Oct 2010), `https://aclanthology.org/D10-1040`
8. Graf, C., Degen, J., Hawkins, R.X.D., Goodman, N.D.: Animal, dog, or Dalmatian? Level of abstraction in nominal referring expressions. In: Proceedings of the 38th Annual Conference of the Cognitive Science Society (2016)
9. Grice, H.P.: Logic and conversation. In: Cole, P., Morgan, J.L. (eds.) Syntax and Semantics, Vol 3: Speech Acts, pp. 41–58. Academic Press, New York (1975)
10. Gualdoni, E.: Object Naming: From Lexical Systems to Language Use and Back. Ph.D. thesis, Universitat Pompeu Fabra (2024)
11. Horn, L.R.: Toward a new taxonomy for pragmatic inference: Q-based and R-based implicature. In: Schiffrin, D. (ed.) Meaning, form, and use in context: Linguistic applications, pp. 11–42. Georgetown University Press (1984)
12. Jolicoeur, P., Gluck, M.A., Kosslyn, S.M.: Pictures and names: Making the connection. Cognitive Psychology **16**, 243–275 (1984)
13. Koolen, R., Goudbeek, M., Krahmer, E.: The effect of scene variation on the redundant use of color in definite reference. Cognitive Science **37**, 395–411 (2013). https://doi.org/10.1111/cogs.12019
14. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., Bernstein, M.S., Fei-Fei, L.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision **123**, 32–73 (2017). https://doi.org//10.1007/s11263-016-0981-7

15. Mitchell, M.: Typicality and object reference. In: Proceedings of the 35th Annual Meeting of the Cognitive Science Society. pp. 3062–3067 (2013)
16. Mitchell, M., van Deemter, K., Reiter, E.: Natural reference to objects in a visual domain. In: Proceedings of INLG 2010 (2010)
17. Mädebach, A., Torubarova, E., Gualdoni, E., Boleda, G.: Effects of task and visual context on referring expressions using natural scenes. In: Proceedings of the 44th Annual Meeting of the Cognitive Science Society (2022), retrieved from https://escholarship.org/uc/item/7cs7204s
18. Pechmann, T.: Incremental speech production and referential overspecification. Linguistics **27**(1), 89–110 (1989)
19. Rohde, H., Seyfarth, S., Clark, B., Jaeger, G., Kaufmann, S.: Communicating with cost-based implicature: A game-theoretic approach to ambiguity. In: Proceedings of the 16th Workshop on the Semantics and Pragmatics of Dialogue - Full Papers. SEMDIAL, Paris, France (Sep 2012), `http://semdial.org/anthology/Z12-Rohde_semdial_0015.pdf`
20. Rubio-Fernandez, P.: How redundant are redundant colour adjectives? An efficiency-based analysis of color overspecification. Frontiers in Psychology **7**(153), 1–15 (2016)
21. Rubio-Fernandez, P.: Redundant color words are more efficient than shorter descriptions (2019), psyArXiv 10.31234/osf.io/gbpt3
22. Rubio-Fernandez, P.: Color discriminability makes over-specification efficient: Theoretical analysis and empirical evidence. Humanities and Social Sciences Communications **8**(147), 1–15 (2021)
23. Sedivy, J.C.: Pragmatic versus form-based accounts of referential contrast: Evidence for effects of informativity expectations. Journal of Psycholinguistic Research **32**, 3–23 (2003)
24. Silberer, C., Zarrieß, S., Boleda, G.: Object naming in language and vision: A survey and a new dataset. In: Proceedings of the Twelfth Language Resources and Evaluation Conference. pp. 5792–5801. European Language Resources Association, Marseille, France (May 2020), `https://aclanthology.org/2020.lrec-1.710`
25. Viethen, J., Dale, R.: The use of spatial relations in referring expression generation. In: Proceedings of the Fifth International Natural Language Generation Conference. pp. 59–67. INLG '08, Association for Computational Linguistics, USA (2008)
26. Westerbeek, H., Koolen, R., Maes, A.: Stored object knowledge and the production of referring expressions: The case of color typicality. Frontiers in Psychology **6** (2015). https://doi.org/10.3389/fpsyg.2015.00935

A Tag lists

A.1 Fine-grained tags

The following are the fine-grained tags we used to initially classify the data, along with an illustrative example. Files containing the lemmas used for each of these tag categories can be found at `github.com/peter-sutton/referring_expressions`.

Tag	Example
`DetN`	*the bird*
`DetNN`	*the tennis player*
`action`	*the flying bird*
`state_verb`	*the brightly lit restaurant*
`with`	*the plane without a logo*
`text`	*the plane with klm visible on tail*
`part_of_object`	*the bird with outstretched wings*
`position`	*the plane next to the vehicles*
`colour`	*the red car*
`size`	*the biggest bird*
`shape`	*the train with a circular window on the front*
`orientation`	*the horizontal plane*
`age`	*the baby cow*
`material`	*the leather chair*
`other_adj`	*the single clock tower*
`visual_pattern`	*the ornate curtain*

A.2 Grouped tags

Fine-grained tags were then grouped into tags for super-categories as shown below.

`visual` = {`colour`, `material`, `shape`, `size`, `visual_pattern`, `age`, `text`}
`pos_orient` = {`position`, `orientation`}
`eventuality` = {`action`, `state_verb`}
`part_other_obj` = {`part_or_object`}

Fine-grained tags not included in super-category tags. Data tagged with the fine-grained tags `DetN` and `DetNN` were discarded for the analysis of the *syntax necessary* condition, since these referring expressions clearly did not pick out one of the two competing images. Data tagged with the fine-grained tag `with` were not included in the `part_or_object` grouped tag, since they were subsumed by the latter tag. Finally, data tagged with the fine-grained tag `other_adjective`

were discarded for our analyses, since the number of instance of this tag was very low, as as the name suggests, it was more of a wastepaper basket for modifiers over which we could not find generalisations. Instances included *single* in *the single clock tower*, and *full* in *The man with a full beard.*