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

Marina Bolea, Petter R. Sutton, Louise McNally
Universitat Pompeu Fabra
{marina.bolea, peterroger.sutton, louise.mcnally}@upf.edu

Upshot. We present an analysis of data from a large-scale reference dataset that provides both new evidence for the salience of entity *parts* and also finer-grained information about the relative salience of different types of visual and other features when speakers must differentiate target from distractor entities.

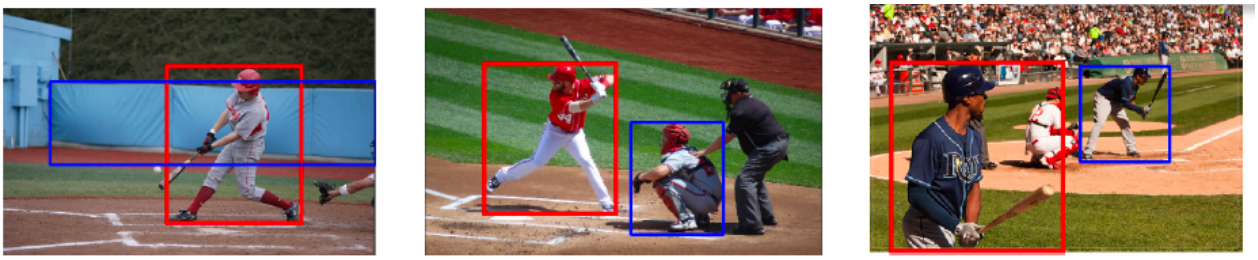
Introduction. For over a decade (see Golland et al. 2010, Rohde et al. 2012, Degen, Franke, et al. 2013, for early examples), controlled experimental studies have established that speakers adjust the referential expressions they use to take into account other candidate referents in a given context. Broadly, 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 principles of efficient communication. But if "efficient" is defined strictly in terms of minimizing unnecessary entailments (e.g., uttering "Dalmatian" or "blue hat" when "dog" or "hat" would suffice), some studies (Graf et al. 2016, Rubio-Fernandez 2016) have revealed that speakers are *not* always maximally efficient. 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., Rubio-Fernandez 2019, Degen, Hawkins, et al. 2020). However, the empirical basis upon which to theorize about speaker behavior remains limited, and, due to experimental design considerations, previous studies arguably have had limited ecological validity.

This study. We carried out a quantitative and qualitative study of modifier use collected in a portion of Mädebach et al.'s 2022 dataset of referring expression production in a referent-identification task involving naturalistic visual images, described below. These images naturally afforded participants a wider range of features for referring expression choice than have previous studies, allowing us to look at preferences speakers manifest *when they have a choice*. In addition to using visual features such as color or size, speakers appealed to features such as position or orientation, as well as events, processes or states in which referents participate and, especially, reference to salient parts of the referent or other objects in the vicinity. Our goal, in addition to gaining insight into speaker preferences in formulating referring expressions, was also to uncover dependencies between these features, including whether some kinds of information exclude others (e.g., whether pre-nominal visual information tends to exclude other pre- and post-nominal information), and whether certain kinds of information tended to be accompanied by additional descriptive support (e.g., whether reference to a salient part or other object is likely to be combined with other modifiers).

Dataset. Mädebach et al., 2022 built on earlier work by Silberer et al., 2020. Via Amazon Mechanical Turk, they presented 97 participants with 72 images drawn from Visual Genome (Krishna et al. 2017), each of which included a target entity (in a red bounding box), and a distractor entity (in a blue box). The images were equally divided among three conditions: *no-competitor*, *lexicon-sufficient*, or *syntax-necessary* (see Fig. 1). Here we limit discussion to the latter two conditions, which were the only ones we analyzed. In the lexicon sufficient condition, there was a lexical item that could distinguish the target from the distractor (e.g., *batter* in Fig. 1 (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. 1 (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 Mädebach et al. 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. We separately analyzed these items (825 in total) as well as the expressions produced in the syntax necessary condition (1439 in total).

We tagged each expression with the type of information used to modify the head noun, using four labels: **visual** information, including color, material, shape, size, age, visual patterns and text (e.g., *the leather chair*, *the BJ83 taxi*); information regarding the **position or orientation** of the target (e.g., *the horizontal plane*, *the chair against the wall*); descriptions of actions or states (**eventuality**, e.g., *the man doing a trick*, *the brightly lit restaurant*) and 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*). These four tags 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, we also noted whether the modifier appeared to left or to the right of the head noun.



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

Figure 1: Sample images from Mädebach et al., 2022

Results. In the ***syntax-necessary*** condition, visual information most frequently modified the noun (see Fig. 2). Referring to a part of the target or to another object within the bounding box was the second most frequent strategy. By splitting the tags into left and 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** 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 feature when syntactic modification was used, with the gap between 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** was the most frequent attribute, followed by position/orientation, visual and eventuality, with the differences between feature frequency being weaker than in the *syntax-necessa*ry condition.

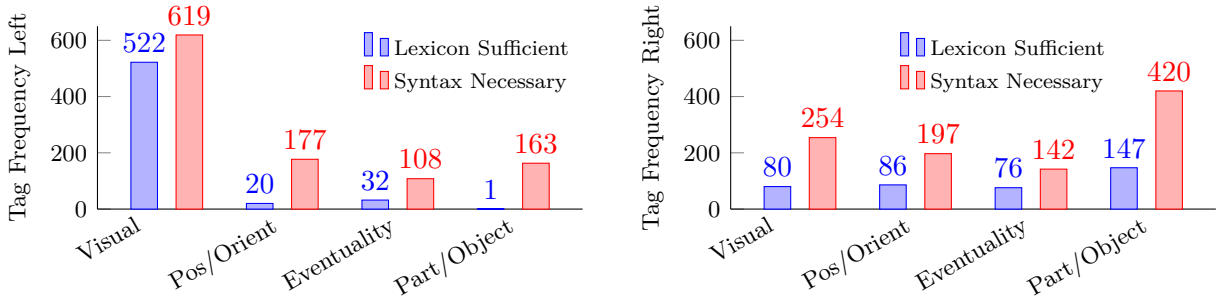


Figure 2: Tag frequency by position in the NP for the *lexicon-sufficient* and *syntax-necessary* conditions, for visual, position/orientation, eventuality, and part or other object tags.

A Bayesian analysis of the two conditions 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* and $\approx 12\%$ of cases in the *syntax-necessary* condition (left in Fig. 3). In contrast, part/other object information on the right correlates with the use of other information on the right (right in Fig. 3). The effect is stronger in the *syntax-necessary* condition ($\approx 73\%$ of cases) than in the *lexicon-sufficient* condition ($\approx 53\%$ of cases).

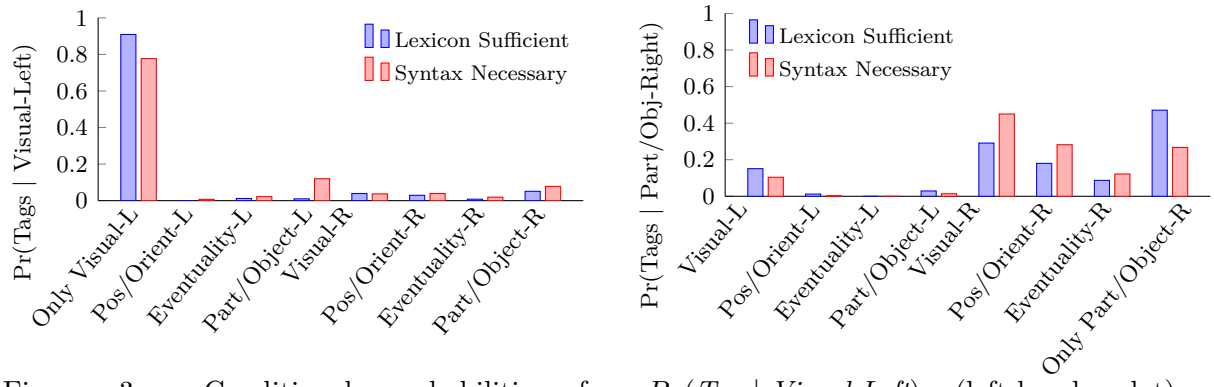


Figure 3: Conditional probabilities for $Pr(Tag \mid Visual\ Left)$ (left-hand plot) and $Pr(Tag \mid Part/Other\ object\ Right)$ (right-hand plot) for the *lexicon-sufficient* and *syntax-necessary* conditions. NB: Aside from Only Visual/Only Part, the categories conditioned upon Visual-Left and Part/Object-Right are not disjoint, since tags can co-occur.

Discussion. Our results echo previous findings, e.g. regarding the salience of color 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 Mitchell et al., 2010). 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. 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.

References

Degen, Judith, Michael Franke, and Gerhard Jäger (2013). "Cost-based pragmatic inference about referential expressions". In: *Proceedings of the Thirty-Fifth Annual Conference of the Cognitive Science Society*, pp. 376–381 (cit. on p. 1).

Degen, Judith, Robert X. D. Hawkins, et al. (2020). "When redundancy is useful: A Bayesian approach to "overinformative" referring expressions". In: *Psychological Review* 127.4, pp. 591–621. DOI: https://doi.org/10.1037/rev0000186 (cit. on p. 1).

Golland, Dave, Percy Liang, and Dan Klein (Oct. 2010). "A Game-Theoretic Approach to Generating Spatial Descriptions". In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Cambridge, MA: Association for Computational Linguistics, pp. 410–419. URL: https://aclanthology.org/D10-1040 (cit. on p. 1).

Graf, Caroline et al. (2016). "Animal, dog, or Dalmatian? Level of abstraction in nominal referring expressions". In: *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (cit. on p. 1).

Krishna, Ranjay et al. (2017). "Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations". In: *International Journal of Computer Vision* 123, pp. 32–73. DOI: https://doi.org/10.1007/s11263-016-0981-7 (cit. on p. 1).

Mädebach, Andreas et al. (2022). "Effects of task and visual context on referring expressions using natural scenes". In: *Proceedings of the 44th Annual Meeting of the Cognitive Science Society*. Retrieved from https://escholarship.org/uc/item/7cs7204s (cit. on pp. 1, 2).

Mitchell, M., K. van Deemter, and E. Reiter (2010). "Natural Reference to Objects in a Visual Domain". In: *Proceedings of INLG 2010* (cit. on p. 3).

Rohde, Hannah et al. (Sept. 2012). "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*. Paris, France: SEMDIAL. URL: http://semdial.org/anthology/Z12-Rohde_semdial_0015.pdf (cit. on p. 1).

Rubio-Fernandez, Paula (2016). "How redundant are redundant colour adjectives? An efficiency-based analysis of color overspecification". In: *Frontiers in Psychology* 7.153, pp. 1–15 (cit. on p. 1).

— (2019). "Redundant color words are more efficient than shorter descriptions". PsyArXiv. DOI: 10.31234/osf.io/gbpt3 (cit. on p. 1).

Silberer, Carina, Sina Zarrieß, and Gemma Boleda (May 2020). "Object Naming in Language and Vision: A Survey and a New Dataset". English. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 5792–5801. ISBN: 979-10-95546-34-4. URL: https://aclanthology.org/2020.lrec-1.710 (cit. on p. 1).