

# Humans Meet Models on Object Naming: A New Dataset and Analysis

Carina Silberer<sup>†‡</sup>    Sina Zarriß<sup>◊</sup>    Matthijs Westera<sup>‡</sup>    Gemma Boleda<sup>‡</sup>  
†University of Stuttgart    ◊Friedrich Schiller University Jena    ‡Universitat Pompeu Fabra  
CarinaSilberer@gmail.com    sina.zarriess@uni-jena.de  
matthijs.westera@gmail.com    gemma.boleda@upf.edu

## Abstract

We release ManyNames v2 (MN v2), a verified version of an object naming dataset that contains dozens of valid names per object for 25K images. We analyze issues in the data collection method originally employed, standard in Language & Vision (L&V), and find that the main source of noise in the data comes from simulating a naming context solely from an image with a target object marked with a bounding box, which causes subjects to sometimes disagree regarding which object is the target. We also find that both the degree of this uncertainty in the original data and the amount of true naming variation in MN v2 differs substantially across object domains. We use MN v2 to analyze a popular L&V model and demonstrate its effectiveness on the task of object naming. However, our fine-grained analysis reveals that what appears to be human-like model behavior is not stable across domains, e.g., the model confuses people and clothing objects much more frequently than humans do. We also find that standard evaluations underestimate the actual effectiveness of the naming model: on the single-label names of the original dataset (Visual Genome), it obtains  $-27\%$  accuracy points than on MN v2, that includes all valid object names.

## 1 Introduction

Research on object naming (Ordonez et al., 2016; Graf et al., 2016; Eisape et al., 2020), such as the linguistic analysis of the naming behavior of humans and computational models, requires natural and reliable object naming data. Such data should contain naturally occurring *naming variation*, i.e., that the same object can very often be called by different names. For instance, a duck can be called *duck*, *bird*, *animal*, etc. At the same time, in cases where a naming dataset provides multiple names for a given object, it is important to verify that this apparent naming variation is not a mere consequence of, for instance, lexical errors or different people naming different objects in the same scene. In this work, we assess the factors that affect the collection of object naming data in the typical Language & Vision (L&V) setup, and test whether these factors impact the accuracy or apparent accuracy of a L&V object labeling model.

Existing work in L&V has relied on data collection methods that prompt natural language users to freely talk about or refer to particular objects in an image (Kazemzadeh et al., 2014; Yu et al., 2016; Silberer et al., 2020). Common to these methods is the use of images as a proxy for the actual context of language use (i.e., the real world), and the use of bounding boxes drawn onto the image to indicate the target object that is to be named. These two common aspects, essentially simulating real-world linguistic reference, have enabled large-scale data collection leveraging existing Computer Vision datasets. However, both aspects also introduce potential confounding factors, such as *referential uncertainty*, where a bounding box may fail to uniquely identify the target object, or *visual uncertainty*, i.e., an object may be more difficult to recognize from a still image than in a real-world scenario.

In this work we analyze, after expansion with new meta-annotations, an existing dataset that was created using the method just described, namely ManyNames (Silberer et al., 2020).<sup>1</sup> It builds upon object annotations in Visual Genome (Krishna et al., 2016), and provides up to 36 name annotations for each of

This work is licensed under a Creative Commons Attribution 4.0 International License. License details: <http://creativecommons.org/licenses/by/4.0>.

<sup>1</sup>The ManyNames data is available at <https://github.com/amore-upf/manynames>

25K objects in images, where the objects belong to 7 different domains (e.g., animals, people, food). To crowdsource the names, Silberer et al. presented subjects with an image and asked them to freely name the target object that was marked by a bounding box. The resulting dataset appeared to exhibit naming variation, with an average of roughly 3 distinct names per object, but, as mentioned in Silberer et al., the presence of naming errors in the data prevent one from drawing conclusions about naming behavior, in humans and in computational models. Indeed, upon manual inspection we found errors of the kinds we mentioned above: some subjects named a different object from the one highlighted by the bounding box, or intended to name the correct object but used a clearly incorrect name (see Figure 1 for examples).

Our first contribution is to assess the factors that affect the collection of valid naming data in the typical L&V setup. To this end, we collect and analyze verification annotations for ManyNames via crowdsourcing. Our analysis shows a) that the predominant type of error in ManyNames is a referential mistake (where subjects name an object other than the target); b) that naming variation in ManyNames is still substantial after correcting for errors and uncertainty, with objects having an average of 2.2 distinct names; and c) that there are differences between domains in the confounding factors, and in the degree of genuine naming variation. We publish a verified version of ManyNames (called ManyNames v2) that we derived through our analysis, to serve as a reliable, verified object naming dataset for future work.<sup>1</sup>

As our second contribution, we assess if the confounders identified by means of our analysis impact the accuracy of a L&V object labeling model. We introduce a diagnostic evaluation based on ManyNames v2. It moves beyond the single-label setup that is common in computer vision and L&V while avoiding the aforementioned confounds. That is, the verification annotations of ManyNames v2 enable us to establish whether a name predicted by the model coincides with the most frequent human response, a less frequent but still valid name, or a mistake (and, in the latter case, of which kind). To the best of our knowledge, to date there exists no work in L&V that performs a systematic investigation of the kind we do here for object naming. We showcase the potential of this evaluation method by analyzing the performance of a popular L&V model, Bottom-up (Anderson et al., 2018). We show a) that a single-label evaluation greatly underestimates the naming capabilities of Bottom-Up, which actually come close to our estimated human upper bound; and b) that, however, its effectiveness varies across domains. We furthermore find c) that the model is confused by the same aspects that confuse humans (as categorized in ManyNames v2), but not always to the same degree, which suggests that the gap between humans and models regarding actual human object naming is still larger than the overall accuracy on its own would suggest.

## 2 Related Work

**Object Naming** Objects are members of many categories and can be called by many names (e.g., a duck can be called *duck*, *bird*, *animal*, etc.). The task of *object naming*—generating a formally correct and also appropriate, *natural* name for an object—has been studied in psycholinguistics and L&V research, and is related to object recognition tasks in Computer Vision. We briefly discuss each area.

Psycholinguistic studies have traditionally focused on object categories instead of individual objects, e.g., the category duck as opposed to a particular duck, and have typically used prototypical or schematic depictions to represent a category (e.g., Rossion and Pourtois, 2004). Such studies have found that humans, when naturally naming an object, have a preference towards a particular name, defined as the *entry-level name* (Rosch et al., 1976; Rosch, 1978; Jolicoeur et al., 1984).

In contrast, work in L&V is mostly concerned with naming particular object instances situated in naturalistic images. In such a setting, naming preferences are more nuanced: humans may prefer different names for instances of the same class (e.g., Fig. 1a-b), and even disagree in their choice for the same instance (Graf et al., 2016; Silberer et al., 2020).

While Graf et al. (2016) modeled object naming in a controlled setting, Ordonez et al. (2016) and Mathews et al. (2015) used ImageNet data (Deng et al., 2009), where images show more realistic, yet still isolated objects annotated with WordNet synsets (Fellbaum, 1998). Zariwaz and Schlangen (2017) train a classification-based model for names produced in referring expressions in the RefCOCO data (Yu et al., 2016), where objects are situated in complex scenes and names might be affected by context. In contrast to our work, existing approaches did not have access to name annotations from many different annotators

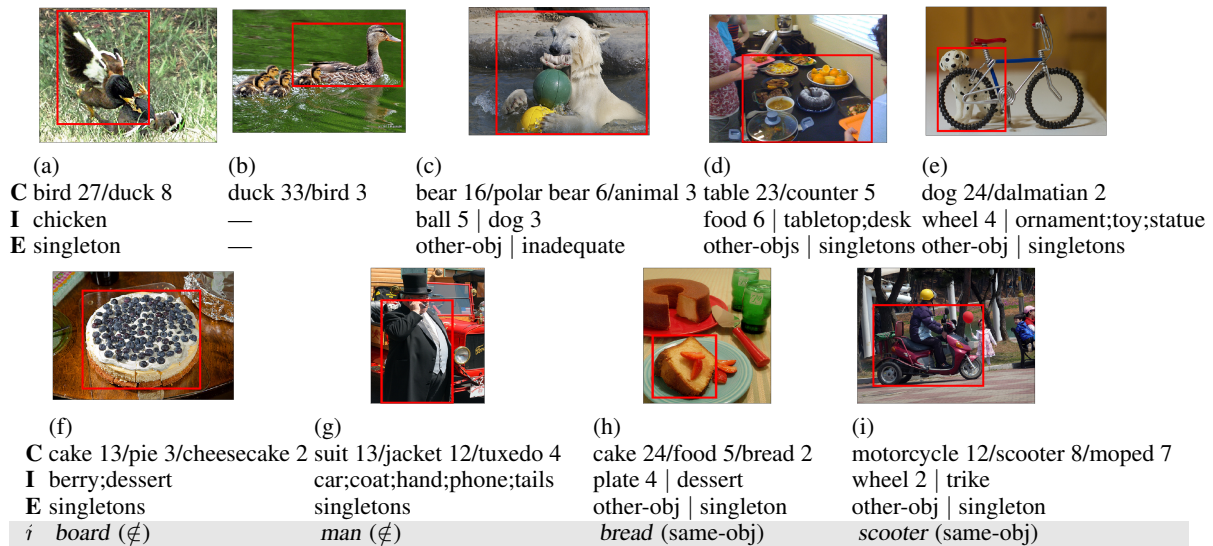


Figure 1: Example instances of ManyNames: target objects, marked by the red boxes, together with their response sets (names and their counts, i.e., the number of MN annotators giving the name; no counts are shown for singletons (where *count*= 1), which are separated by ;). Sets are separated into (C)orrect (adequate) and (I)ncorrect names, and (E)rror types of incorrect names. Names of different objects are separated by |. The gray-shaded row contains predictions of Anderson et al.’s (2018) model ( $\hat{n}$ ).

(as in ManyNames, Silberer et al. 2020) or to the verification data we present here (i.e., ManyNames v2), which enable a more fine-grained evaluation. Beyond diagnostic testing, clean, natural and variable object naming data could also support model development: e.g., Peterson et al. (2019) suggest that object classifiers trained on distributions over object labels are more robust and generalize better.

Finally, object recognition tasks in Computer Vision have been agnostic about phenomena intrinsic to human object naming, such as naming variation. In these tasks the goal is usually to assign a single supposed ground-truth label from a pre-defined vocabulary, varying from 20 categories (Everingham et al., 2015) to a few thousand synsets (Russakovsky et al., 2015) or words (Kuznetsova et al., 2020). Although previous work considered effects of object/image characteristics on model effectiveness (Hoiem et al., 2012), a comparison to a human upper bound of the kind we conduct has thus far been missing.

**Uncertainty in Bounding Box Annotations** Bounding boxes are by far the most common way to annotate objects in images, despite the fact that they are well-known to cause problems for annotation efficiency and quality (Papadopoulos et al., 2017; Kuznetsova et al., 2018; Hoiem et al., 2012). In Computer Vision, the standard protocol to *obtain* box annotations is the one established by Russakovsky et al. (2015), which asks annotators to draw a tight box around each object, namely, the smallest box containing all visible parts of the object. Kuznetsova et al. (2018) point out, however, that this criterion can still trigger uncertainty (e.g., is *water part of a fountain?*). Moreover, the protocol itself has proven difficult to strictly adhere to. In the collection of Visual Genome (Krishna et al., 2016), annotators produced region descriptions and corresponding bounding boxes for objects. Even though these boxes were additionally verified and linked in a separate verification round, Silberer et al. (2020) observe that the resulting annotations still do not fully comply with the bounding box annotation protocol (e.g., the same object can have multiple distinct boxes), in some cases leading to referential uncertainty.

Other L&V datasets have relied on predefined box annotations to prompt new annotators or speakers with a specific target (Kazemzadeh et al., 2014; Yu et al., 2016; De Vries et al., 2017). These approaches assume that verification of the data is ensured by the interactive protocol, in which a second participant is listening and must identify the target object. However, it does not overcome the problem that the initial speaker may already face referential uncertainty in their box prompt (e.g., *water* or *fountain*). To the best of our knowledge, no prior work in L&V systematically investigates this, as we do here for object naming.

### 3 Free Object Naming: Verification Collection and Analysis of ManyNames v2

#### 3.1 Obtaining Consistent Response Sets from ManyNames v1

Silberer et al. (2020) created the dataset of object names, ManyNames, by collecting on Amazon Mechanical Turk (AMT) approximately 36 names each for 25k target objects  $o$ . Each object  $o_j$  was presented visually by a picture containing a single bounding box  $b_j$  to delineate the target object, see Figure 1 for examples. The ManyNames annotations are structured as follows: for a box  $b_i$  that delineates an object  $o_i$  in image  $i \in \mathcal{I}$ , we have a **response set**  $R_i := \{(n_1, p_1), \dots, (n_k, p_k)\}$  of  $k$  name–frequency pairs. Unless otherwise stated, by **name–object pairs** we refer to name *types* given an individual object (i.e., there are  $k$  names). For each object, there are  $M$  **name tokens**, with  $M := \sum_{l=1}^k \text{count}(n_l)$ . For each name  $n_j$  in the response set,  $p_j$  is the relative frequency of that name in the response set:  $p_j := \frac{\text{count}(n_j)}{M}$ . This means that  $p$  can be interpreted as the estimated probability distribution over names  $n_j$  for box  $b_i$ . Let  $n_{top}$  denote the preferred or **top name**, i.e., the most frequent name the AMT workers gave for  $b_i$ . Let  $n_{alt}$  denote the remaining, less preferred or **alternative names**.<sup>2</sup>

Due to visual and referential uncertainty in the images (see Section 1), or plain annotation errors, ManyNames 19 provides no guarantee that a name in a response set is in fact an adequate name for the target object, or that two names which occur in the same response set were even intended for the same object. For instance, annotators might have entered an inadequate name, such as *dog* for the bear in Figure 1c; or named different objects in the box, such as the bear or the ball in the same image. Our goal is to obtain **consistent** response sets  $R_i$  which only contain adequate names  $n_j \in R_i$  for the same object  $o_i$  that is delineated by  $b_i$ . Hence, we must identify errors in ManyNames along two dimensions: (i) the **adequacy** which verifies each individual name–box pair  $(n_j, b_i)$  and (ii) **same-object/other-object** which verifies each name–name–box triple  $(n_l, n_j, b_i)$  with respect to object identity, i.e., whether the annotators who provided  $n_l$  and  $n_j$  for the box  $b_i$  likely intended to name the same object. Fig. 1d shows that, because of referential uncertainty of boxes, adequacy alone is not enough to identify consistent response sets: *food* was judged fully adequate given the box, but also as naming a different object from the (likewise adequate) top name *table*. Given both adequacy and same-object annotations, we will be able to compute a consistent response set for each box as follows: given its top name  $n_{top}$ , exclude an alternative name  $n_{alt}$  from the response set if it does not refer to the same object, or if it has low adequacy.

#### 3.2 Collection of the Verification Annotations: ADEQUACY, INADEQUACYTYPE, SAMEOBJECT

We recruited annotators via crowdsourcing using Amazon Mechanical Turk (AMT)<sup>3</sup>. Workers were asked to conduct exactly the two types of annotations mentioned above: adequacy and same-object.<sup>4</sup> For adequacy, workers could choose between “perfectly adequate” (which we encoded with score 1 for the analysis below), “there may be a (slight) inadequacy” (score 0.5) and “totally inadequate” (score 0). In addition, if a worker selected a slight or total inadequacy they had to specify the type of inadequacy from a pre-defined list (based on our prior data inspection): *referential* (paraphrased as “named object not tightly in bounding box”; e.g., *bear–ball* in Fig. 1c), *visual* recognition (“named object mistaken for something it’s not”; as in *bear–dog*), *linguistic* (such as *dear* for *deer*), and “something else” (*other*).<sup>5</sup>

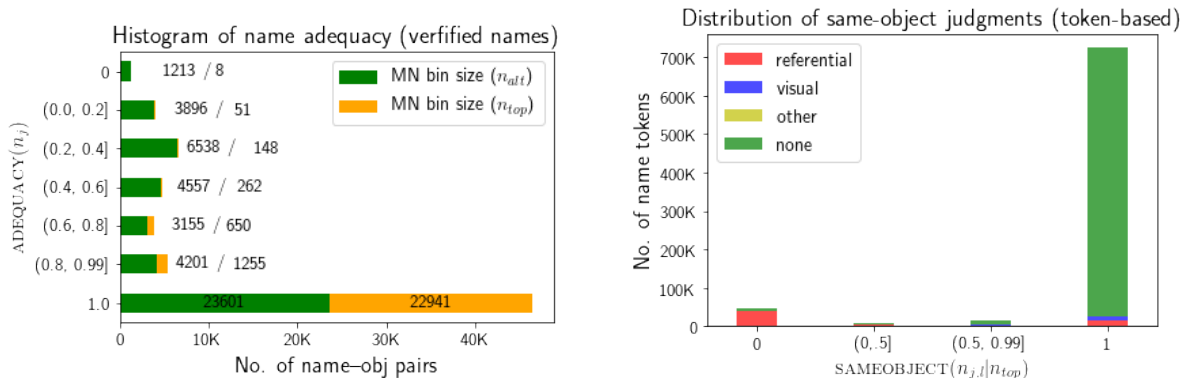
In this way we collected verification annotations for the entire ManyNames data, except for objects where all annotators gave the same name, which we assumed were reliable, and names with a count of 1, which we assumed were unreliable. The remaining 19, 427 images had on average around 4 names per object, totalling 69, 356 name–object pairs to be verified. Each image, with the target object marked by a box, was presented to 3 workers, along with its ManyNames response set (minus names that were given only once). In total, 255 participants contributed annotations.

<sup>2</sup>Note that ManyNames uses only images and boxes from Visual Genome, not the names; the name that Visual Genome originally assigned to an object  $o_i$  is not necessarily an element of the ManyNames response set  $R_i$ .

<sup>3</sup><https://mturk.com>

<sup>4</sup>For adequacy, we provided the definition that “a name is adequate if there is an object in the image, whose visible parts are tightly circumscribed by the red bounding box, that one could reasonably call by that name”.

<sup>5</sup>Below, we report the results for *linguistic* mistakes as part of the *other* causes, because it was very infrequent (less than 1% of all name–object pairs), possibly because Silberer et al. (2020) did automatic spelling correction.



(a) Distribution of ADEQUACY scores for name–object pairs, divided into 7 bins, each bin further subdivided into top names (orange) and alternative names (green) for each object.

(b) Distribution of SAMEOBJECT scores for pairs of top names and their alternative names (weighed by the name counts in the original ManyNames response set), divided into 4 (uneven) bins, each bin further subdivided by inadequacy type (colors).

Figure 2: Distribution of ADEQUACY and SAMEOBJECT scores in ManyNames.

Given the annotations, we compute for each name  $n_j$  its mean adequacy score among its 3 collected annotations, which we denote by  $\text{ADEQUACY}(n_j)$ . Let  $\text{INADEQUACYTYPE}(t|n_j)$  denote the score of name  $n_j$  for each inadequacy type  $t$ , computed as the proportion of annotators who judged  $n_j$  to have inadequacy type  $t$  (representing  $t$  as ‘none’ when the annotator selected ‘perfectly adequate’). Likewise, let  $\text{SAMEOBJECT}(n_j|n_l)$  denote the proportion of annotators who judged  $n_j$  to be intended to name the same object as  $n_l$ .

### 3.3 Analysis: Errors, their Causes, and Effects

Figure 2a shows the histogram (in 7 bins) of ADEQUACY (mean adequacy scores), each bin further subdivided into top names ( $n_{top}$ , orange) and alternative names ( $n_{alt}$ , green) for each object. The figure shows that most names are fully or largely adequate, with 68% of name–object pairs judged as ‘perfectly adequate’ by all respective annotators (ADEQUACY = 1). In contrast, only few name–object pairs were judged as ‘completely inadequate’ by all annotators (0.03% of top names and 2.4% of alternative names). There is a clear division between the top names  $n_{top}$  in orange, which are overwhelmingly perfectly adequate, and the alternative names  $n_{alt}$  in green, of which half (52%) are judged as perfectly adequate and the other half spread along the full range of mean adequacy scores. The former reflects that a name that was produced by many different humans (for 96% of all ManyNames objects, the  $n_{top}$  was given by at least 10 people) is very likely to be a perfectly adequate name for the target object. Overall, most names are fully or largely adequate. As we will show next, most inadequacies correspond to referential issues.

**Mostly Referential Errors** We will next show that most inadequacies correspond to referential issues, and that these names, however, have a low response frequency, by looking at both, name–object pair types (as in Fig. 2a) and tokens. Table 1 (row ManyNames v1) represents our verification annotations for the full ManyNames dataset. It shows the distribution of inadequacy types across name–object pairs, jointly considering slight and total inadequacies; with an average INADEQUACYTYPE of 21%, most naming issues are indeed referential (i.e., cases where names do not exactly correspond to the object delimited by the bounding box). However, Table 1 reports name–object pair types; if we consider tokens (individual responses) instead, we find referential errors in 7% of the cases. That is, names with referential issues have a low response frequency. A less prominent but still noticeable issue are visual recognition errors (4%), while other errors occur rarely. We will explain the remaining rows of Table 1 in Section 3.4.

The fact that most INADEQUACYTYPES are referential has the effect that the most prevalent cause of noise in the data is subjects naming other objects than the one that, according to most annotators, was the one marked by the box. Figure 2b illustrates this with the SAMEOBJECT scores (i.e., proportion of annotators judging two names to be the same object); it shows the *token*-based distribution of the

SAMEOBJECT scores between all name responses and their corresponding top names  $n_{top}$ , divided into 4 (uneven) bins, each bin further subdivided by INADEQUACYTYPE (colors). We consider *name tokens*—the name that was given by each individual ManyNames annotator for a given object—, not name types—the name that was given by at least 2 annotators (recall that we discarded names with count 1)—in order to better reflect the distribution of actual naming behavior. To obtain the token-based distribution, we multiplied the SAMEOBJECT scores by the name counts in the original ManyNames response sets (see Appendix A for details). Note the strong agreement in the same-object judgments: 91% of all name tokens were judged to name the same object as  $n_{top}$  by all our annotators (SAMEOBJECT = 1), and 6% to not (SAMEOBJECT = 0). Only in 3% of cases (the middle two bins) did our annotators disagree on whether a given token and  $n_{top}$  are co-referring. Also note the expected correspondence between the error types and the SAMEOBJECT judgments: when SAMEOBJECT = 0, most pairs are deemed to be referential errors. This shows that the task of deciding whether two ManyNames annotators, who entered different names, were likely intending to name the same object elicits robust judgments.

**Discussion of Causes** We have shown that there is a non-negligible number of cases in which subjects named an object different from the target. Part of the reason may be annotators not doing their task faithfully; recall that the initial data collection was a generation task in which it is difficult to put quality control mechanisms. However, the high amount of perfectly adequate names suggests that this is unlikely to be the sole or even main cause, pointing perhaps to the experimental setup itself as a culprit: the real-world context of naming behavior is imperfectly approximated by an image and a potentially ambiguous box. Qualitative analysis is consistent with this hypothesis. On the one extreme, we find plain annotator errors, such as *ball* for a bear in Figure 1c, and on the other, cases of genuine bounding box ambiguity, as in Figure 1d, where it is not possible to determine whether the box marks the table or the food it contains.<sup>6</sup> Most cases fall in between these extremes, with effects like object salience clearly playing a role. For instance, in Figure 1e, the box marks the dog, but the wheel occupies almost the whole box and is more visually salient, even occluding the target object. Most people correctly identified the dog, but four named the wheel instead. These effects are partially domain-dependent, as we will discuss in Section 3.5.

### 3.4 Defining ManyNames v2

For the analyses that follow, here and in Section 4, we define a new version of ManyNames containing only consistent response sets (see Sect. 3.1). Names in a consistent response set must name the same object as the top name in the set, which we define in terms of SAMEOBJECT > 0, and must be sufficiently adequate, which we define as ADEQUACY > 0.4. Thus, consistent response sets are obtained by removing from the original ManyNames (henceforth ManyNames v1) all names that do not meet these two criteria. The set of excluded names corresponds to row 3 of Table 1, and row 4 shows the dataset obtained by excluding them, which we refer to (and publicly release) as ManyNames v2. Accordant with the fact that most issues are referential, there is a large overlap between the two criteria: only 10% of the removed names are discarded by the addition of the adequacy threshold.

We chose the thresholds of our two criteria based on the foregoing analysis, and with the aim of excluding clear errors while leaving room for borderline cases or genuine disagreements between annotators. One of our research interests being naming variation, we did not want to exclude a potential variant just because, e.g., a single annotator considered it inadequate. However, for different tasks it may make sense to choose these thresholds differently, and to facilitate this we will publicly release the raw verification annotations in addition to the consistent response sets as we computed them. Our analysis below (Sect. 3.5) will focus on the consistent response sets as defined, i.e., ManyNames v2. For model analysis in Section 4 we will also rely on the different inadequacy types.

### 3.5 Analysis of Naming Variation per Domain: ManyNames v1 vs. v2

The consistent response sets of ManyNames v2 give us a more reliable empirical window on genuine naming variation, i.e., the existence of multiple adequate names for the same object. Table 2 compares

---

<sup>6</sup>Using segmentation masks instead of boxes could partially remedy referential ambiguity, but Computer Vision datasets providing masks are much smaller than those with boxes, and some issues (like highly overlapping objects) would still remain.

	inadequacy type (%)				# name-obj pairs
	ref.	vis.	oth.	none	
ManyNames v1	21	4	1	73	72,476
SAMEOBJECT( $n_{alt} n_{top}$ )=0	87	2	3	8	14,338
$\dots \cup$ ADEQUACY( $n_{alt}$ ) $\leq 0.4$	81	7	4	8	15,948
ManyNames v2	4	4	1	92	56,528

Table 1: Distribution of inadequacy types for ManyNames (first row) and different subsets: cases where SAMEOBJECT score is 0 (second row), cases where either that holds or ADEQUACY  $\leq 0.4$  (third row), and ManyNames v2, obtained from ManyNames v1 by removing all cases in the third row.

domain	ManyNames v1		ManyNames v2	
	N	% $n_{top}$ $\pm$ std	N	% $n_{top}$ $\pm$ std
all	2.9	75.2 $\pm$ 21.9	2.2	80.2 $\pm$ 20.7
people	4.3	59.0 $\pm$ 20.4	3.3	65.1 $\pm$ 21.8
clothing	3.2	70.1 $\pm$ 18.5	2.4	76.7 $\pm$ 18.1
home	3.1	72.6 $\pm$ 20.7	2.1	81.2 $\pm$ 19.1
buildings	3.0	74.7 $\pm$ 20.7	2.1	82.7 $\pm$ 19.3
food	2.9	76.4 $\pm$ 20.7	2.4	79.7 $\pm$ 19.3
vehicles	2.4	76.6 $\pm$ 19.8	2.1	78.9 $\pm$ 19.6
animals/plants	1.5	94.5 $\pm$ 12.1	1.3	95.4 $\pm$ 11.4

Table 2: Naming variation in ManyNames v1 vs. v2, with mean number of distinct names per object (columns N) and mean percentage of  $n_{top}$  responses (% $n_{top}$ ).

(apparent) naming variation in ManyNames v1 and v2, overall and by domain (rows), by listing both the mean number of names per object (columns  $N$ ) and the mean percentage of original ManyNames annotators who entered the top name. The table shows a 25% reduction in mean number of names per object (from 2.9 for ManyNames v1 to 2.2 for ManyNames v2 on average), and an increase in the percentage of entered names being the top name (75% to 80%). Thus, inevitably, noise in the data led to overestimating naming variation; however, even after noise removal, substantial variation remains. This suggests that free object naming cannot be adequately modeled with a single-label approach, as is common in Computer Vision—we return to this in Section 4.

ManyNames v1 on its own already suggests that variation is domain-dependent, a picture that is maintained in ManyNames v2: people trigger the most variation (3.3 names on average in ManyNames v2), and animals the least (1.3 names). Moreover, by comparing across domains also the *reduction* in variation of ManyNames v2 compared to v1 we can see that the susceptibility to factors that inflate true variation, primarily referential and visual inadequacies, is domain-dependent too. People, home, and buildings are the most susceptible to these factors ( $-1$  in  $N$  for ManyNames v2); vehicles and animals/plants the least ( $+0.3$  and  $-0.2$ ). In the former, most discarded items involved referential errors or uncertainty (about 85%, in contrast to, e.g., food with 67%), typical example errors being the confusion of clothes and the wearer of them, a background target with a foreground object (e.g., Fig. 1e), or ambiguous bounding boxes (e.g., Fig. 1d). In contrast, the animals/plants domain has the most non-referential inadequacies (43.6%; clothing the least with 3.2%), which are predominantly visual errors, likely reflecting the visual similarity of different types of animals, especially as seen on pictures. In these domains humans seem to have a strong tendency towards the ‘basic-level’ category (e.g., *bear* is preferred over the hyponym *polar bear* and the hypernym *animal*), explaining their lower naming variation. Interestingly, this preference holds even in the face of visual uncertainty, i.e., cases where a hypernym would have been safer, such as *animal* in the case where two subjects entered *goat*, one entered *horse* and another *deer* (for ‘deer’).

Finally, we remark that ensuring consistent response sets (filtering incorrect names) proved the most difficult for the food domain. In ManyNames v2, the food domain has the largest proportion of name tokens (7.6%) for which it is still unclear if they name the same object as the top name (i.e.,  $0 < \text{SAMEOBJECT} < 1$ ), followed by clothing (4.9%). The associated inadequacies are primarily referential, but also visual (e.g., uncertainty about what the picture is showing) and linguistic (terminology). The food domain in particular seems to be susceptible to variation between the annotators in categorization/terminology. For example, *bread* in Fig. 1h survived into ManyNames v2 as an adequate alternative for *cake*, even though it was given by only 2 original ManyNames v1 annotators.

#### 4 Diagnosing Model Effectiveness in Human-Like Object Naming

In Section 3.4 we obtained ManyNames v2, which provides a reliable categorization of object names: the top name, alternative names for the same object (i.e., in the consistent response set), adequate names for



other objects (i.e., outside the consistent response set), and inadequate names of various types. We now use it to define a diagnostic evaluation method for object naming models, one which is more fine-grained than the predominant single-label evaluation. It also allows to assess whether models are affected by the same issues as humans, in particular referential and visual uncertainty.

We apply our evaluation to Bottom-Up (Anderson et al., 2018) as a representative L&V object naming model, which has been widely used for transfer learning in L&V (Lu et al., 2019; Gao et al., 2019; Chen et al., 2019; Cadene et al., 2019; Tan and Bansal, 2019, *inter alia*). In contrast to existing works in Computer Vision research (Hoiem et al., 2012) on diagnosing the effects of object or image characteristics on model performance, ManyNames allows us to compare the model against an upper bound of the human performance in object naming, estimated via the verification annotations.

Our analysis focuses on two questions: First, can an object detector that was trained in a single-label setting (i.e., towards predicting unique ground truth names) account for the naming variation inherent in human object naming behavior? Second, does the model exhibit a similar sensitivity as humans to the interaction between domains (humans, clothes, etc.) as well as the visual characteristics of target objects?

#### 4.1 Experimental setup

As for the diagnostic evaluation method, let  $\hat{n}$  be the name that is predicted by Bottom-Up for a given image with a bounding box indicating a target object  $o_i$  for this image. The method checks whether  $\hat{n}$  is in the object’s consistent response set according to ManyNames v2, further subdividing the positive cases by whether  $\hat{n}$  matches the top name  $n_{top}$  (**correct. $n_{top}$** ) or one of its adequate alternative names (**correct.same-object**). We treat cases where the predicted name  $\hat{n}$  is not in the consistent response set as **incorrect** (although this rests on assumptions, see below), which we subdivide with the help of the verification data we collected, distinguishing cases in which (i)  $\hat{n}$  must have been intended for a different object (i.e.,  $\text{SAMEOBJECT}(\hat{n}|n_{top}) = 0$ , henceforth **incorrect.other-obj**), (ii)  $\hat{n}$  was intended for the target object but is **inadequate** ( $\text{SAMEOBJECT}(\hat{n}|n_{top}) > 0 \wedge \text{ADEQUACY}(\hat{n}) \leq 0.4$ ), (iii) a name that was given by only a single annotator (**incorrect.singleton**), and (iv) a name that did not occur in the ManyNames v1 response set (**incorrect.unobserved**). Our treatment of categories (iii) and (iv) as incorrect is a simplifying assumption to facilitate analysis.<sup>7</sup> Among the incorrect cases we treat (i)-(iii) as human-like errors, because at least one ManyNames v1 annotator produced the name, and (iv) as a non-human-like error, assuming a human would not give that name.

We use the object labeling model Bottom-Up (Anderson et al., 2018)<sup>8</sup>, which builds upon the Faster R-CNN architecture (Ren et al., 2015), and which was initialized with features pre-trained on 1K ImageNet classes (Deng et al., 2009; Russakovsky et al., 2015) with the ResNet-101 classification model (He et al., 2016). The model was originally optimized for a set of 1,600 frequent names in Visual Genome (VG).

For evaluation we use those names in the Bottom-Up vocabulary which also occur among the 7,970 names in ManyNames v2, resulting in a target vocabulary of 1,253 names. We test on the subset of ManyNames images that are included in the VG test split that was used by Anderson et al. (2018), and whose top ManyNames name is covered by the evaluation vocabulary (1,145 images in total). We compare model effectiveness against the human upper bound, which is computed by taking all name tokens of an object’s response set in MN v1 as name predictions of a ‘human model’, and applying our evaluation methodology to them.

#### 4.2 Results

Table 3 shows the main results of Bottom-Up and the human upper bound. Bottom-Up achieves an accuracy of 73.4% on the top names (first column), but when taking all correct names into account its accuracy is +14.5% points higher (87.9%, third column). This shows that the standard single-label evaluation of recognition models in Computer Vision substantially underestimates model effectiveness, punishing models for what are actually valid alternatives. It also illustrates, especially for the evaluation of L&V methods, the importance of taking into account linguistic variation when assessing model

<sup>7</sup>Recall that we excluded singletons from the verification phase (even if some may be correct, see Fig. 1).

<sup>8</sup>We used the code and model available at [github.com/peteanderson80/bottom-up-attention](https://github.com/peteanderson80/bottom-up-attention). The model is trained with an additional output over attributes, but we only use the object/instance class prediction layer.



Model	$n_{top}$	correct		incorrect			
		same-object	all	other-object	inadequate	singleton	unobserved
Human	75.9	15.2	<b>91.1</b>	1.8	3.1	3.9	–
Bottom-Up	73.4	14.5	87.9	2.5	1.5	1.0	7.1

Table 3: Results of the human upper bound vs. Bottom-Up: numbers represent the proportion (in %) of predicted names that fall under each category (see text for details on the splits).

Model	Domain	$n_{top}$	correct		incorrect				total #
			same-object	all	other-object	inadequate	singleton	unobs.	
Human	people	59.6	<u>28.2</u>	87.8	2.1	4.1	5.9	–	224
Bottom-Up	people	70.1	18.8	<b>88.9</b>	1.3	0.0	3.6	6.2	224
Human	clothing	74.2	17.5	<b>91.7</b>	2.4	1.6	4.2	–	97
Bottom-Up	clothing	59.8	16.5	76.3	2.1	0.0	9.3	12.4	97
Human	food	73.0	18.6	<b>91.6</b>	1.0	3.0	4.4	–	98
Bottom-Up	food	69.4	12.2	81.6	2.0	0.0	2.0	14.3	98
Human	buildings	77.3	9.5	86.8	3.1	4.6	5.5	–	50
Bottom-Up	buildings	68.0	14.0	82.0	2.0	2.0	4.0	10.0	50
Human	vehicles	76.5	18.1	94.6	0.8	1.7	2.9	–	182
Bottom-Up	vehicles	68.1	<u>25.3</u>	93.4	0.5	0.0	1.6	4.4	182
Human	home	74.7	12.3	87.0	3.0	5.8	4.1	–	279
Bottom-Up	home	71.3	13.3	84.6	2.9	3.6	1.8	7.2	279
Human	animals_plants	95.1	2.2	97.3	0.4	0.4	1.9	–	215
Bottom-Up	animals_plants	93.5	2.8	96.3	0.0	0.0	0.0	3.7	215

Table 4: Results per domain, again as % of predicted names that fall under each category. Column ‘total #’ gives the number of test instances per domain.

effectiveness on human language in visual scenes (Vedantam et al., 2015; Jedoui et al., 2019). Not shown in the table, we found that evaluating Bottom-Up using the supposed ground-truth name  $n_{vg}$  from Visual Genome (the dataset upon which ManyNames is built), instead of  $n_{top}$  from ManyNames, underestimates model effectiveness even further (down to 61.2% accuracy, not in the table). This demonstrates that many annotations are needed (such as the 36 of ManyNames) for the top name to accurately reflect naming preferences. We refer to Appendix B for a more detailed analysis of this.

As for the human upper bound, Bottom-Up comes close in general (87.9% accuracy, compared to 91.1% for humans), and even has a similar distribution across correct top and correct alternative names of around 74% and 15%, respectively (columns 1 and 2). Among Bottom-Up’s incorrect predictions, almost half are human-like according to our categorization, i.e., cases where Bottom-Up predicted a name for another object, an inadequate name, or a singleton (2.5 + 1.5 + 1.0 vs. 7.1, i.e., 41% vs. 59%). Overall, we conclude that Bottom-Up can accurately simulate human object naming in images, and, like humans, is affected by visual and referential uncertainty caused by the task setup, foremost, by relying on bounding boxes in images to delineate a target object.

However, a closer look per domain, given in Table 4, reveals that confounding factors do affect Bottom-Up and humans differently. Bottom-Up has a higher overall accuracy (third column) than humans on the people domain (88.9% vs. 87.8%), but 15% points worse on the clothing domain (76.3% vs. 91.7%). Qualitative analysis shows that the model is, as humans, quite susceptible to referential issues in the people and clothing domains, but with the effect of learning a bias towards people: it tends to recognize the wearer rather than the clothing item (e.g., Bottom-Up predicted *man* for the clothing item in Fig. 1g). Indeed, 45% of cases in other-object and singletons, and 75% of unobserved cases were due to the model predicting a person instead of a clothing name.

The only other domain in which Bottom-Up falls short against the human bound by quite a margin is food (81.6% accuracy, 10% points lower). Qualitative analysis revealed that 67% of its incorrect name predictions (most of which are in unobserved) are related to referential and/or visual issues. This has

the effect of confusing the depicted object with kitchenware (see the *cake-board* example in Fig. 1f) or with visually similar objects. In some cases the model’s predictions (e.g., *bread* in Fig. 1h) are also controversial for humans and subject to personal differences (see Section 3.5).

With respect to predicted naming variation, we find contrasts in the vehicles and people domains: In vehicles, although the Bottom-up’s overall accuracy is human-like, it has a weaker preference for the top name than humans, favoring alternative names. These alternatives often involve synonymy (e.g., *airplane-plane*) as well as difficult-to-name objects where the top and alternative names seem equally plausible (e.g., Fig. 1i). We find the reverse in the people domain, where Bottom-Up predicts the top name *more* often than humans, i.e., human responses are more varied.

In sum, we have shown that, when taking natural naming variation into account, a representative labeling model performs close to humans, in two respects: its overall accuracy, and its tendency to predict the top name around 74% of the time, alternative correct names around 15%, and incorrect names in the remaining cases. At the same time, we found differences between model and human behavior, in particular in the people, clothing and food domains, where the model exhibits less variation than humans, has learnt a bias towards a competing domain, or generally performs worse. This highlights the importance of fine-grained evaluation on the basis of reliable data that captures both natural naming variation and errors, such as ManyNames v2.


## 5 Conclusions

Modeling how humans use language in the visual world is at the core of L&V research. We have focused on object naming, the choice of a noun (or compound noun) to refer to an object which is marked with a bounding box in a real-world image. This setup is typical of Computer Vision and L&V for tasks such as object classification, referring expression interpretation/generation or visual dialogue. Our findings underline the importance of modeling naming as a phenomenon of its own: A woman, for example, can be named *skier*, *person*, or *woman*, in different images or by different people. At the same time, there are clear preferences about how to name a particular object: overall naming agreement in humans is around 80%. To boost research on object naming, we provide a high-quality naming resource, ManyNames v2, which is a verified version of a previous dataset with 36 names for objects in 25K images.

For dataset collection, our analysis of the verification data strongly supports a collection methodology that elicits names from *many* speakers, in order to capture the variation in possible naming choices, and to reliably estimate the preferred name of an object. It furthermore suggests that referential issues, the main cause of noise introduced through the collection setup, can be greatly reduced by a simple verification step that assumes that the object named by the most frequent response is the target object, and asks subjects to select the alternative names for this object.

For model development and evaluation, both naming variation and its domain dependence need to be taken into account. We have shown that ManyNames provides a very different picture of the performance of a state-of-the-art naming model, compared to a resource that only provides one single gold name per object (Visual Genome). Our analysis also shows that the model’s naming behavior differs from that of humans particularly in the people, clothing and food domains, that is, in domains that are very familiar and relevant in human daily life, and in which we have found that humans exhibit the highest language variation. This is based on a model that provides a single answer per object; future work should seek to do even more justice to language variation, by predicting the whole probability distribution of names for objects (Peterson et al., 2019). We hope that our work will spur further research on object naming and, more generally, how humans use language to talk about the world.

## 6 Acknowledgments

We thank the anonymous reviewers for their comments, and the AMT workers who participated in our crowdsourcing task. This project has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement No715154) and by the Catalan government (SGR 2017 1575). This paper reflects the authors’ view only, and the EU is not responsible for any use that may be made of the information it contains. 

## References

- Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *Proceedings of CVPR*.
- R. Cadene, H. Ben-younes, M. Cord, and N. Thome. 2019. MUREL: Multimodal Relational Reasoning for Visual Question Answering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1989–1998.
- Hui Chen, Guiguang Ding, Zijia Lin, Sicheng Zhao, and Jungong Han. 2019. Cross-Modal Image-Text Retrieval with Semantic Consistency. In *Proceedings of the 27th ACM International Conference on Multimedia, MM '19*, page 1749–1757, New York, NY, USA. Association for Computing Machinery.
- Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 2017. Guesswhat?! visual object discovery through multi-modal dialogue. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5503–5512.
- Jia Deng, W. Dong, Richard Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.
- Tiwalayo N Eisape, Roger Levy, Joshua B Tenenbaum, and Noga Zaslavsky. 2020. Toward human-like object naming in artificial neural systems. In *Proceedings of ICLR 2020 Workshop on Bridging AI and Cognitive Science*.
- Mark Everingham, S. M. Eslami, Luc Gool, Christopher K. Williams, John Winn, and Andrew Zisserman. 2015. The pascal visual object classes challenge: A retrospective. *Int. J. Comput. Vision*, 111(1):98–136, January.
- Christiane Fellbaum. 1998. *WordNet*. Wiley Online Library.
- Peng Gao, Zhengkai Jiang, Haoxuan You, Pan Lu, Steven C. H. Hoi, Xiaogang Wang, and Hongsheng Li. 2019. Dynamic Fusion With Intra- and Inter-Modality Attention Flow for Visual Question Answering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June.
- Caroline Graf, Judith Degen, Robert XD Hawkins, and Noah D Goodman. 2016. Animal, dog, or dalmatian? level of abstraction in nominal referring expressions. In *Proceedings of the 38th annual conference of the Cognitive Science Society*. Cognitive Science Society.
- Kenji Hata, Ranjay Krishna, Li Fei-Fei, and Michael S. Bernstein. 2017. A Glimpse Far into the Future: Understanding Long-Term Crowd Worker Quality. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, CSCW '17*, pages 889–901. Association for Computing Machinery.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. pages 770–778.
- Derek Hoiem, Yodsawalai Chodpathumwan, and Qieyun Dai. 2012. Diagnosing error in object detectors. In Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid, editors, *Proceedings of the 12th European Conference on Computer Vision (ECCV 2012)*.
- Khaled Jedoui, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. 2019. Deep Bayesian Active Learning for Multiple Correct Outputs.
- Pierre Jolicoeur, Mark Gluck, and Stephen Kosslyn. 1984. Pictures and names: Making the connection. *Cognitive psychology*, 16:243–275.
- Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L Berg. 2014. ReferItGame: Referring to Objects in Photographs of Natural Scenes. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, pages 787–798, Doha, Qatar.
- Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. 2016. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations.
- Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Mallocci, Tom Duerig, et al. 2018. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv preprint arXiv:1811.00982*.

- Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Mallocci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. 2020. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*.
- Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23.
- Alexander Mathews, Lexing Xie, and Xuming He. 2015. Choosing Basic-Level Concept Names Using Visual and Language Context. *Applications of Computer Vision*, pages 595–602.
- Vicente Ordonez, Wei Liu, Jia Deng, Yejin Choi, Alexander C. Berg, and Tamara L. Berg. 2016. Learning to Name Objects. *Commun. ACM*, 59(3):108–115, February.
- Dim P Papadopoulos, Jasper RR Uijlings, Frank Keller, and Vittorio Ferrari. 2017. Extreme clicking for efficient object annotation. In *Proceedings of the IEEE international conference on computer vision*, pages 4930–4939.
- Joshua C Peterson, Ruairidh M Battleday, Thomas L Griffiths, and Olga Russakovsky. 2019. Human uncertainty makes classification more robust. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9617–9626.
- Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *NIPS’2015*, pages 91–99.
- Eleanor Rosch, Carolyn B Mervis, Wayne D Gray, David M Johnson, and Penny Boyes-Braem. 1976. Basic objects in natural categories. *Cognitive psychology*, 8(3):382–439.
- Eleanor Rosch. 1978. Principles of Categorization. In Eleanor Rosch and Barbara B. Lloyd, editors, *Cognition and Categorization*, pages 27—48. Lawrence Erlbaum, Hillsdale, N.J., USA.
- Bruno Rossion and Gilles Pourtois. 2004. Revisiting snodgrass and vanderwart’s object pictorial set: The role of surface detail in basic-level object recognition. *Perception*, 33(2):217–236.
- Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252.
- Carina Silberer, Sina Zarrieß, and Gemma Boleda. 2020. Object Naming in Language and Vision: A Survey and a New Dataset. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5792–5801.
- Hao Tan and Mohit Bansal. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, November.
- Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.
- Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg, 2016. *Modeling Context in Referring Expressions*, pages 69–85. Springer International Publishing.
- Sina Zarrieß and David Schlangen. 2017. Obtaining referential word meanings from visual and distributional information: Experiments on object naming. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 243–254, Vancouver, Canada, July. Association for Computational Linguistics.

## A Computing the Token-Based SAMEOBJECT Distribution

To compute the token-based SAMEOBJECT distribution, we use the top names  $n_{top}$  as reference names, and take into account the response frequency of each name type given for an individual object, including  $n_{top}$ , from whose counts we subtract 1. Specifically, we count each occurrence of a name for an object  $i$ , but ignore a single occurrence of the reference name  $n_{top}$  by setting  $\text{count}(n_{i,top}) - 1$ . Overall, for each object, we analyse  $M - 1$  instances of name-object pairs, with  $M - 1 = \sum_{j=1}^{|R_i|} [\text{count}(n_{i,j>0})] - 1$ .

## B Adequacy and Causes of Error in Visual Genome

The Visual Genome dataset (Krishna et al., 2016, VG), upon which ManyNames (MN) is built, used a different name collection methodology, which in principle should prevent referential uncertainty, because, essentially, VG’s AMT workers grounded names in the image by drawing a box, such that they explicitly chose the referent for the name. However, a comparison of MN and VG shows that incorrect name annotations can also be found in VG, despite its collection setup. First, for 27% of the MN objects, the VG name  $n_{vg}$  does not match the top MN name  $n_{top}$ ; i.e.,  $n_{vg}$  is not the name that most people would use for the object. For instance, in Figure 2f, 13 subjects chose *cake*, and only 3 the name given in VG, *pie*. Moreover, a quarter of the non-matching names do not even refer to the same object as  $n_{top}$ , and half of them are inadequate. We refer to Hata et al. (2017) for a thorough analysis of worker quality.