Object Naming in Language and Vision: A Survey and a New Dataset

Carina Silberer, Sina Zarrieß, Gemma Boleda
Universitat Pompeu Fabra, University of Jena, Universitat Pompeu Fabra
Barcelona (Spain), Jena (Germany), Barcelona (Spain)
sina.zarriess@uni-jena.de
{carina.silberer, gemma.boleda}@upf.edu

Abstract

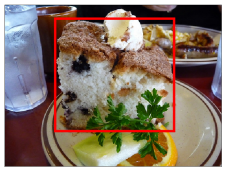
People choose particular names for objects, such as *dog* or *puppy* for a given dog. Object naming has been studied in Psycholinguistics, but has received relatively little attention in Computational Linguistics. We review resources from Language and Vision that could be used to study object naming on a large scale, discuss their shortcomings, and create a new dataset that affords more opportunities for analysis and modeling. Our dataset, ManyNames, provides 36 name annotations for each of 25K objects in images selected from VisualGenome. We highlight the challenges involved and provide a preliminary analysis of the ManyNames data showing that there is a high level of agreement in naming, on average. At the same time, the average number of name types associated with an object is much higher in our dataset than in existing corpora for Language and Vision, such that ManyNames provides a rich resource for studying phenomena like hierarchical variation (*chihuahua* vs. *dog*), which has been discussed a lot in the theoretical literature, and other less well studied phenomena like cross-classification (*cake* vs. *dessert*).

Keywords: object naming, language and vision, computer vision

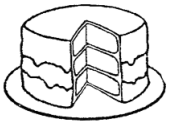
1. Introduction

Generally, research in Language & Vision (L&V) is interested in modeling how speakers naturally talk about visual objects and scenes, in contrast to the fixed image labeling schemes used in Computer Vision. Data collections in L&V are typically set up as free annotation tasks, where subjects are free to produce whatever word or utterance they consider most suitable for the given task (e.g. image description, reference to objects, visual dialogue), which naturally results in linguistic variation. For this reason, large-scale data collections in L&V usually provide a certain amount of parallel annotations for the same entity from different annotators (Fang et al., 2015; Devlin et al., 2015; Kazemzadeh et al., 2014; Mao et al., 2015; De Vries et al., 2017).

Compared to experimental work on perception and language grounding, however, recent data collections in L&V capture a rather limited amount of inter-speaker variation. For instance, picture naming norms used in Psycholinguistics record naming responses for hundreds of speakers for the same object (Snodgrass and Vanderwart, 1980; Rossion and Pourtois, 2004), whereas captioning or referring expression data sets typically provide less than 10 annotations per entity (Devlin et al., 2015; Kazemzadeh et al., 2014; Mao et al., 2015). Consequently, a reliable assessment of inter-annotator agreement, speaker preferences, and a deeper linguistic analysis of the observed variation is mostly not possible with available data collections in L&V. At the same time, these datasets have much potential for research on these topics, as they provide more realistic images of real-world objects and scenes than the idealized drawings used in picture naming norms (see Figure 1), as well as a wider coverage of object categories. A more systematic understanding of the factors influencing the choice of object names could inform theoretical work on language grounding and pragmatics (Rohde et al., 2012; Graf et al., 2016), as well as the design of models and architectures, as e.g. (Lazaridou et al., 2015; Ordonez et al., 2016; Zhao et al., 2017).



cake (53), food (19), bread (8), burger (6), dessert (6), snacks (3), muffin (3), pastry (3)



cake (83)

Figure 1: Names for a cake object in ManyNames (left) and in Snodgrass's Naming Norms (right), percentages of responses in parentheses.

In this paper, we present ManyNames, a dataset with substantial amounts of object names per object for real-world images. We start by surveying existing resources in L&V that provide object names: RefCOCO (Yu et al., 2016), a collection of referring expressions, Flickr30k Entities (Plummer et al., 2015), which provides region-to-phrase linkings for Flickr 30K captions (Young et al., 2014), and VisualGenome (Krishna et al., 2016), which features extensive region-level annotations. We highlight their potential contributions to the study of object naming, and also argue that the low number of annotations per item prevents reliable assessment or linguistic analysis of object-specific preferences and naming variation.

To address this shortcoming, we contribute a new dataset, ManyNames, that contains 36 crowd-sourced names for 25K instances from VisualGenome.[1] The number of annotations per object available in ManyNames is considerably larger than in recent data collections in L&V.

The trends we identify in the dataset are illustrated in Figure 1 (left): Our data reveals clear naming preferences (in the example, 53% of the annotators prefer the so-called basic-level name *cake*, see Section 2.1. for further explanation) and also rich variation (the remaining annotators prefer other

[1] The dataset will be available upon publication.

options like *food, dessert, bread*) that is not restricted to taxonomic relations studied in previous work on naming (Ordonez et al., 2016; Graf et al., 2016): while *food* is in a taxonomic relation to *cake* (it is a hypernym), *dessert* highlights a different facet of the object.

2. Background

2.1. Object Naming as a Linguistic Phenomenon

The act of naming an object amounts to that of picking out a nominal to be employed to refer to it (e.g., "the *dog*", "the white *dog* to the left"). Since an object is simultaneously a member of multiple categories (e.g., a young beagle belongs to the categories DOG, BEAGLE, ANIMAL, PUPPY, PET, etc.), all the various names that lexicalize these constitute a valid alternative, meaning that the same object can be called by different names (Brown, 1958; Murphy, 2004).

Seminal work by Rosch et al. (1976) inspired a taxonomic view of object naming, in which names exhibit a preferred level of specificity or abstraction called the "entry-level" (Jolicoeur, 1984). *//g: Make sure that discussion reflects the fact that entry-level refers to a *taxonomy of concepts/categories*, not to names.//* This typically corresponds to an intermediate level of specificity (basic level, e.g., *bird*, *car*), as opposed to more generic (super-ordinate, e.g., *animal*, *vehicle*) or more specific categories (sub-ordinate, e.g., *sparrow*, *convertible*). However, less prototypical members of basic level categories tend to be instead identified with sub-ordinate categories (e.g., a penguin is typically called *penguin* and not *bird*; Jolicoeur (1984)).

While the traditional notion of entry-level categories suggests that objects tend to be named by a *single* preferred concept, research on pragmatics has found that speakers adapt their naming choices to the context and, hence, are flexible with respect to the chosen level of specificity (Olson, 1970; Rohde et al., 2012; Graf et al., 2016). For example, in presence of more than one dog, the name *dog* is ambiguous and a sub-ordinate category (e.g., *rottweiler*, *beagle*) is potentially preferred by speakers. The effect of such distractor objects on the production of referring expressions has been looked at a lot in the language generation community (Krahmer and Van Deemter, 2012), though not specifically for object naming. We believe that our new dataset provides an interesting resource for tackling this question.

The purely taxonomic view on naming has also been criticized in work on object organization, which found that many objects of our daily lifes are part of multiple category systems at the same time (Ross and Murphy, 1999; Shafto et al., 2011). This *cross-classification* occurs, for instance, with food categories which can be taxonomy-based (e.g. *meat*, *vegetable*) or script-based (e.g. *breakfast*, *snack*). We provide tentative evidence that cross-classification is indeed relevant for naming variation, and that the taxonomic axis is not the most frequent source of variation in our data.

2.2. Picture Naming in Cognitive Science

An important experimental paradigm in work on human vision and categorization is picture naming, where subjects have to say or write down the first name that comes to mind when looking at a picture of (typically) a line drawing depicting a prototypical instance of a category (Snodgrass and Vanderwart, 1980; Rossion and Pourtois, 2004), see Figure 1. Subjects reach very high agreement in this task (Rossion and Pourtois, 2004), i.e. for a given object, there is a clear tendency towards a certain name across all speakers. The resulting naming norms are useful for studying various cognitive processes (Humphreys et al., 1988). Our task is inspired by picture naming, but uses real-world images showing objects in context.

2.3. Object Recognition in Computer Vision

In Computer Vision, object recognition is often modeled as a classification task where state-of-the-art systems localize and classify objects into thousands of different categories (Szegedy et al., 2015; Russakovsky et al., 2015). Current recognition benchmarks use labels and images from the ImageNet (Deng et al., 2009) ontology, and typically assume a single ground-truth label. The construction of ImageNet was set up as a two-stage procedure: (i) images for given categories in the ontology where automatically collected by querying search engines, (ii) crowd-workers then verified whether each candidate image is an instance of the given category. Other data collection efforts for object labels also used a predefined vocabulary and asked annotators to mark all instances of these categories in a set of images (Lin et al., 2014; Kuznetsova et al., 2018). Recently, Pont-Tuset et al. (2019) have argued for annotation of object labels using free form text though here this free vocabulary is then mapped to a set of underlying classes. Thus, even though object recognition benchmarks do provide images of objects and categories, they generally do not provide what we are interested in in this work, namely natural names of objects.

2.4. Object Naming in L&V

Previous work in L&V has collected and used data sets where annotators produced free and natural utterances for a given image. Moreover, these data sets typically record utterances that are more complex than a single word, such as image captions (Fang et al., 2015; Devlin et al., 2015; Bernardi et al., 2016), referring expressions (Kazemzadeh et al., 2014; Mao et al., 2015; Yu et al., 2016), visual dialogues (Das et al., 2017; De Vries et al., 2017) or image paragraphs (Krause et al., 2017). While object names occur in all of these data sets, they are not necessarily marked up and linked to the corresponding image regions. The overview in Section 3. will discuss corpora where the grounding of names to regions for objects is given, as in the case of VisualGenome (Krishna et al., 2016), or where it can be easily derived, as in the case of referring expressions.

Our new collection, ManyNames, focusses on object names in isolation and is substantially more controlled than common L&V data sets. This controlled collection procedure allowed us to elicit many annotations for the same object from different annotators, resulting in a data set that is amenable to studying variation and preferences in naming systematically and on a large scale.

3. Object Names in Existing L&V resources

We identified three previously existing resources that can be of use for analysis and modeling of object naming: RefCOCO (and a variant, RefCOCO+), Flickr30k Entities, and

	RefCOCO/+	Flickr30kE	VG	VGmn	MN
# objects	50.000	243.801	3.781.232	25.315	25.315
naming vocab size	5.004	10.423	105.441	1.061	7.970
av. annotations/object	2.8	2.3	1.7	7.2	35.3
% objects with n types > 1	0.7	0.3	0.02	0.05	0.9
av. types/object	1.9	1.4	1	1.1	5.7

Table 1: Overview statistics for different data sets containing object naming data. VGmn shows statistics for the subset of VG that overlaps with our ManyNames dataset.

Visual Genome. Table 1 summarizes their main characteristics and compares them to our dataset (last two columns; see Section 4.). As the table shows, previous datasets provide between one and three annotations per object, which, we believe, is not enough to assess naming behavior for individual objects and which motivates our data collection. In the following, we will look at their characteristics in more detail and work out requirements for a dataset that is suitable for a large-scale study of object naming.

3.1. RefCOCO and RefCOCO+

Both RefCOCO and RefCOCO+ (Yu et al., 2016) use the ReferIt (Kazemzadeh et al., 2014) game for collecting referring expressions (RE) for natural objects in real-world images, and are built on top of MS COCO (Lin et al., 2014), a dataset of images of natural scenes of 91 common object categories (e.g., DOG, PIZZA, CHAIR). The REs were collected via crowdsourcing in a two-player reference game designed to obtain REs uniquely referring to the target object. Specifically, a director and a matcher are presented with an image, and the director produces a RE for an outlined target object. The matcher must click on the object she thinks the RE refers to. REs in RefCOCO/+ were collected under the constraints that (i) all images contain at least two objects of the same category (80 COCO categories), which results in longer and more complex REs than just the object name, and (ii) in RefCOCO+ the players cannot use location words, urging them to refer to the appearance of objects.
Table 1 shows that the multiple annotations (2.84 on average) actually contain a considerable amount of variation in naming (almost 2 different names on average per object). However, the small number of annotations per object does not allow for a reliable assessment of speaker agreement.
RefCOCO has been used to model and examine the effect of context on referring expression generation in general (Yu et al., 2016), though this work did not look at object names specifically. A controlled analysis of the effect of context on choice in naming, as for instance in (Graf et al., 2016), would require substantial further data annotation as not all objects of an image are annotated with REs and corresponding categories. Hence, so-called distractor objects (Krahmer and Van Deemter, 2012) and their names cannot be analyzed systematically. Also, while in RefCOCO the elicited names can assumed to be natural, it is unclear how the additional constraints in RefCOCO+ impact on the naturalness of object naming. Finally, the set of categories in MSCOCO included is quite small (80 COCO categories). To sum up, RefCOCO is suitable for generally modeling referring expressions in context for a restricted set of categories, but less appropriate for analyzing object naming at a large scale.

3.2. Flickr30k Entities

The Flickr30k Entities dataset (Plummer et al., 2015) augments Flickr30k, a dataset of 30k images and five sentence-level captions for each of the images, with region-level descriptions extracted from the captions. Specifically, mentions of the same entities across the five captions of an image are linked to the bounding boxes of the objects they refer to. This dataset has three main differences with respect to RefCOCO/+: (i) the entity mentions were obtained via an image description task (captioning), as opposed to a referential task; (ii) the images and the production of entity mentions were not subject to any constraints; (iii) a much wider range of categories are covered (cf. the number of objects and the vocabulary size in Table 1). Moreover, although no exhaustive annotations of the images are available, the dataset does contain information for the most salient objects in the image, as they are typically mentioned in the captions. The number of annotations per object, 2.3 annotations, is comparable to RefCOCO. This dataset is suitable to analyze object naming in descriptions, for a quite large set of categories (although, again, not enough annotations are available to analyze image-specific naming data).

3.3. Visual Genome

VisualGenome (VG, Krishna et al. (2016)) is one of the most densely and richly annotated resources currently available in L&V; here, we focus on aspects immediately relevant to object naming. VG aims at providing a full set of descriptions of the scenes which images depict in order to spur complete scene understanding. The data collection followed a complex procedure, involving many different rounds of annotation. The first round of the procedure, and the basic backbone for the further rounds, is a collection of region-based descriptions: workers were asked to describe regions in the image and draw boxes around the corresponding area in the image (for examples, see Figure 2).
In a second, independent round (involving new workers), annotators were asked to process the region descriptions by (i) marking the object names contained in the region description, and (ii) drawing a tight box around the corresponding region. As different region descriptions can potentially mention the same objects, each worker was shown a list of previously marked objects and encouraged to select an existing object rather than annotating a new one.
One of the main advantages of VG are its size, with 3.8 million objects (108K images) as opposed to 50K and 243K for the other two datasets, and its category coverage, with a

vocabulary of object names of 105K compared to 5K/10K. Another is the fact that it in principle provides exhaustive annotations of objects in the image, often with several region descriptions and possibly object names per object. This should make it easier to identify factors intervening in naming choices, and to model contextual aspects that may affect them, than in the case of RefCOCO.

However, there is a crucial pitfall: As Figure 2 shows, there is only a partial linking of objects that are mentioned across different region descriptions; for instance, the first, second, and fifth object IDs in the figure actually correspond to the same object. Moreover, the regions for the beak of the object (third and fourth object IDs) overlap with those of the bird. This means that the identity of objects cannot be established based on the annotation, which severely limits the usefulness of the data to analyze naming. For instance, even though there is a different name (*vulture*) for *bird* object in Figure 2, the annotation suggests that *bird* is the only available name. The relatively low number of annotations per objects in VisualGenome (1.7 on average) shown in Table 1 and the very small number of objects that have more than one name associated with it (5%) seem to be an effect of this partial linking problem. We experimented with filtering and merging bounding boxes based on overlap, but this would introduce substantial noise into the data (e.g. truly overlapping objects).

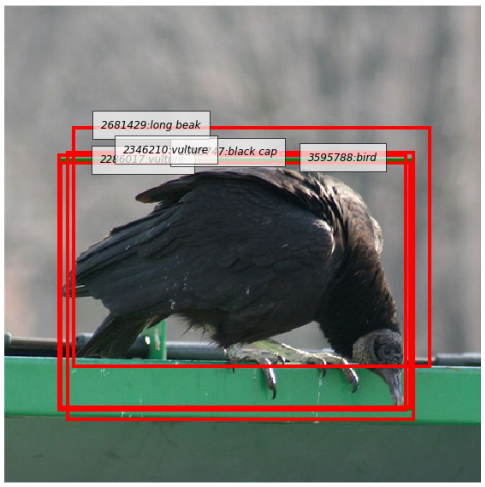
Table 1 also shows the statistics for the subset of those VG objects that we selected for ManyNames and, here, we find a considerably higher average of annotations per object (7). We think that this might be an effect of our category selection procedure explained in Section 4.. However, interestingly, the portion of objects that have different names associated with them is still extremely small. Note that in contrast, even though RefCOCO has much less annotations per object, there are many objects with different names (70%).

3.4. Discussion

While some existing resources do provide naming data for objects in context, they do not provide *enough* data to systematically assess how variable or stable object naming really is. The RefCOCO data (and to some extent the Flickr30k data) suggests that for most objects there is more than one available name, but it is unclear which name most speakers would prefer or whether there is such a preferred name at all. The VG data, to the contrary, seems to indicate that the vast majority of objects should only be associated with a single name, but it is difficult to estimate to what extent this finding results from problems with annotation (partial linking). This shows that to be able to analyze object naming in more detail, we simply need naming data from many subjects for the same objects. Also, dense annotations of images can be beneficial to analyze the factors affecting naming (e.g., the category or salience of other objects), and how these impact the modeling of natural language in L&V. These are the motivations for our dataset, ManyNames, and for building it on top of VG, as discussed next.

4. A New Dataset: ManyNames

We take data from VisualGenome (VG) because its dense annotations of images can be beneficial to analyze the factors affecting naming (e.g., the category or salience of other objects), and how these impact the modeling of natural language in L&V. VG suits our purpose of collecting names for naturalistic instances of common objects, as it has images of varying complexity, with close-ups as well as images with many objects. Moreover, its object names are linked to WordNet synsets (Fellbaum, 1998), which affords analysis possibilities that we will exploit in Section 5. below. Note that, as common in Computer Vision, objects in VG images are localized as bounding boxes, as shown in Figure 1 (left).[2]



object id	linked region descriptions
3595788	the **bird** is black in color, nose of the **bird**, a **bird** relaxing in stand, small white beak of **bird**, large black talon of **bird**, a **bird** on a green pole, a green bar under **bird**, black **bird** on green rail, small black eye of **bird**
2286017	large black **vulture** on fence, a vulture on bar
2385747	small white beak of **bird**
2681429	a semi **long beak**
2346210	a black and gray **vulture**

Figure 2: Bounding boxes, names and region descriptions for an object in VisualGenome

4.1. Sampling of Instances

We selected images from seven domains: ANIMALS_PLANTS, BUILDINGS, CLOTHING, FOOD, HOME, PEOPLE, VEHICLES. They are all based on McRae et al.'s (McRae et al., 2005) feature norms, a dataset widely used in Psycholinguistics that comprises common objects of different categories, except for PEOPLE, which we added because it is a very frequent category in VG and a very prominent category for humans.

Within each domain, we aimed at collecting instances at different taxonomic levels to cover a wide range of phenomena, but this is not straightforward because ontological taxonomies do not align well with the lexicon (for instance, *dog* and *cow* are both mammals, but *dog* has many more

[2]We use image and object interchangeably in the following, since we only selected one target object per image (i.e., each object and image in VG is chosen at most once).

Domain	Collection synsets
animals_plants	ungulate$_1$ (2037), horse$_1$ (833), feline$_1$ (763), dog$_1$ (688) bird$_1$ (389), flower$_1$ (44), rodent$_1$ (27), insect$_1$ (12), fish$_1$ (11)
buildings	house$_1$ (364), bridge$_1$ (297), shelter$_1$ (169), restaurant$_1$ (58), outbuilding$_1$ (31), hotel$_1$ (19), housing$_1$ (17), place_of_worship$_1$ (12)
clothing	shirt$_1$ (968), overgarment$_1$ (786), dress$_1$ (199), headdress$_1$ (135), neckwear$_1$ (65), robe$_1$ (27), glove$_2$ (7), footwear$_1$ (5)
food	dish$_2$ (812), baked_goods$_1$ (770), foodstuff$_2$ (280), vegetable$_1$ (48), edible_fruit$_1$ (42), beverage$_1$ (23)
home	furnishing$_2$ (5,355), vessel$_3$ (525), kitchen_utensil$_1$ (132), crockery$_1$ (92), cutlery$_2$ (82), tool$_1$ (72), lamp$_1$ (34)
people	woman$_1$ (1768), man$_1$ (1167), male_child$_1$ (853), athlete$_1$ (396), child$_1$ (333), creator$_2$ (11), professional$_1$ (5)
vehicles	aircraft$_1$ (1208), train$_1$ (957), car$_1$ (727), motorcycle$_1$ (564), truck$_1$ (559), boat$_1$ (499), ship$_1$ (38)

Table 2: Overview of the ManyNames dataset: Synset nodes for each domain (subscript indicates synset number; number of instances in parentheses).

vehicles	food	animals_plants	home	buildings	people	clothing
train (954)	pizza (518)	giraffe (915)	bed (888)	house (340)	boy (853)	shirt (904)
car (642)	cake (261)	horse (822)	bench (714)	bridge (274)	man (806)	jacket (451)
plane (485)	bread (186)	cat (754)	table (687)	dugout (91)	woman (766)	coat (267)
airplane (479)	sandwich (153)	dog (654)	desk (672)	tent (53)	girl (650)	dress (190)
motorcycle (466)	bun (143)	zebra (461)	counter (516)	restaurant (33)	lady (342)	hat (77)

Table 3: Overview of the ManyNames dataset: Top 5 VG names for each domain (number of instances in parentheses).

common subcategories), and most domains are not organized in a clear taxonomy in the first place (e.g. HOME). Instead, we defined a set of 52 synsets (listed in Table 2) that we used to collect object instances from VG, as follows. First, to create our synset set, we chose those VG synsets that match or subsume the object classes in the McRae norms, and cover different names in VG. For example, VG instances subsumed by McRae's *dog* were named *dog*, *beagle*, *greyhound*, *bulldog*, etc., while McRae's *duck*, *goose*, or *gull* did not have name variants in VG, so we kept *dog* and *bird* (which subsumes *duck*, *goose*, or *gull*) as collection synsets. We then retrieved all VG images depicting an object whose name matches a word in these collection synsets or in those subsumed by them. We refer to the names obtained as *seeds* (450 in total). We did not consider objects with names in plural form, with parts-of-speech other than nouns[3], or that were multi-word expressions (e.g., *pink bird*). We further only considered objects whose bounding box covered $20 - 90\%$ of the image area. Because of the Zipfian distribution of names, and to balance the collection, we sampled instances depending on the size of the seeds: up to 500 instances for seeds with up to 800 objects, and up to 1000 instances for larger seeds. This yielded a dataset of $31,093$ instances, which was further pruned during annotation, as explained next. Table 3 shows the top 5 VG names in each of the domains.

4.2. Elicitation Procedure

To elicit object names, we set up a crowdsourcing task on Amazon Mechanical Turk (AMT). In initial pilot studies, we found object identification via bounding boxes to be problematic. In some cases, the bounding box was not clear; in others, AMT workers named objects that were more salient than the one signaled by the box (for instance, for a box around a jacket, the man wearing it). We took special care of minimizing this issue in two ways: Specifying the instructions such that workers pay close attention to what object is being indicated in the box, and pruning images with unclear boxes or occluded objects via an initial collection round in which we allowed workers to mark such cases. Figure 3 shows the task instructions for this first round, in which 9 workers annotated each image.

After the first round, and based on the opt-out annotation, we kept images that met all the following conditions (thresholds were estimated via manual inspection): (i) they were not marked as occluded by any subject; (ii) "Bounding box is unclear" was marked at most twice; (iii) at most 17% of elicited names were in plural form (to remove cases where the bounding box contains several objects); (iv) the most frequent elicited name is of the same domain as the VG name. This yielded $25,596$ images (we discarded $5,497$). We then did 3 more collection rounds, obtaining a total of 36 images per object. Figure 4 shows the instructions for these rounds; they were accompanied by a FAQ solving common issues. We shuffled the set of images per task between rounds, and workers could only participate in one round, to avoid workers annotating an instance more than once. Overall 841 workers took part in the data elicitation, with a median of 261 instances (range $= [9, 17K]$) per worker.

5. Analysis

As shown in Table 1 above, ManyNames gathers many more names per object than previous datasets: 35.3 on average, compared to 1-7. It also contains the most variability, since objects have on average 5.7 names (compared to 1-1.9). Figure 5 shows some example datapoints of ManyNames with high and low name agreement. ManyNames shows high potential use for studying the degree of inter-subject naming agreement, and what factors influence variation. Data analysis shows that object identification remains an issue

[3]We obtained tags with CoreNLP (Manning et al., 2014).

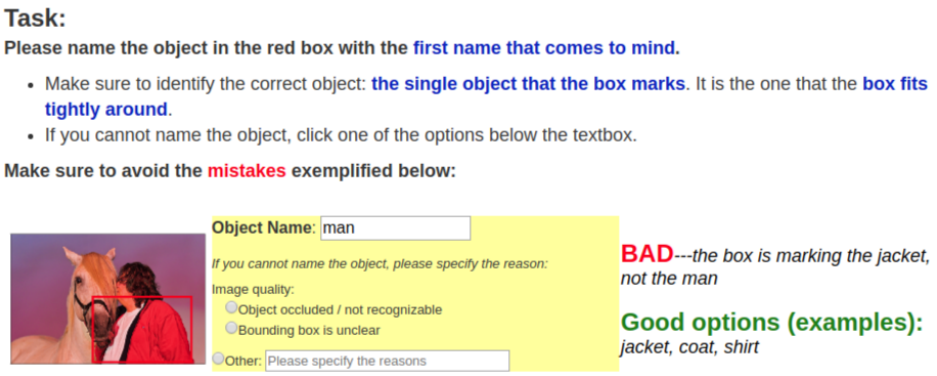


Figure 3: Instructions for AMT annotators for the first round (whole instructions showed more examples, see Figure 4).

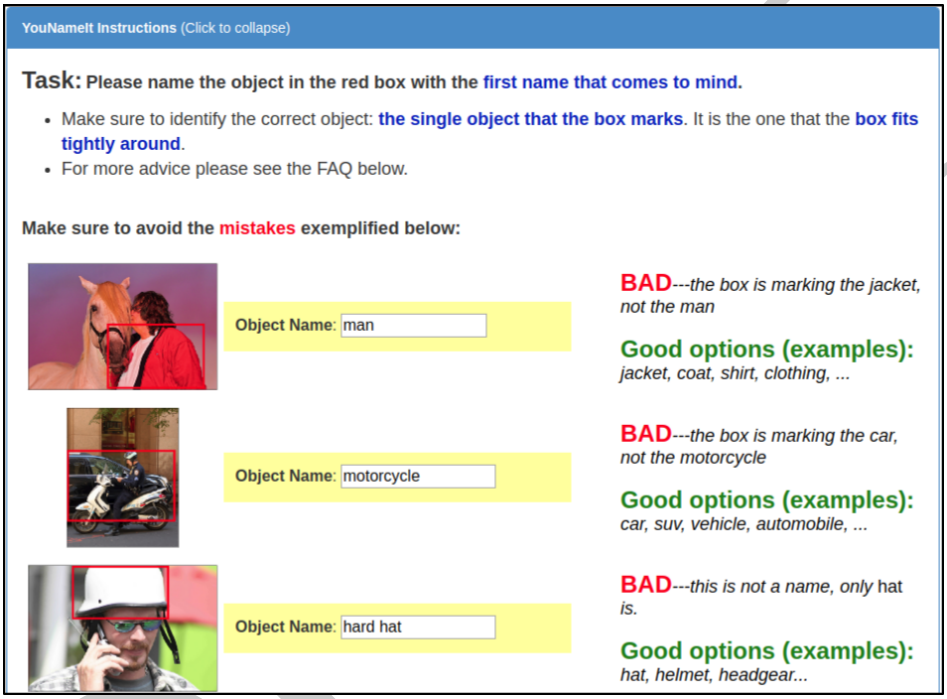
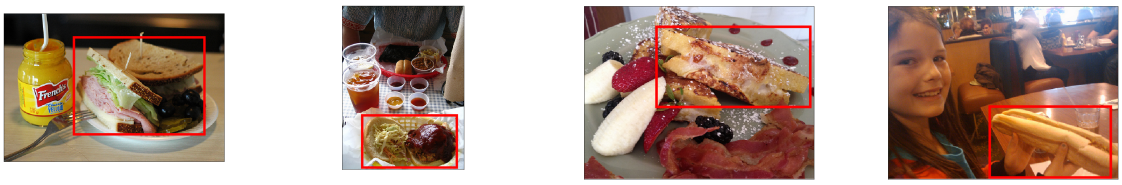


Figure 4: Instructions for AMT annotators for rounds 2 to 4.

in our data, though: Despite our care in filtering out objects that are occluded or have unclear bounding boxes (see Section 4.), we still find many examples where annotators identified different objects for the same bounding box. Typically, workers named an adjacent object or one supported by the target object (such as *toy/book* instead of *bed* in Fig. 5, image K), or a part of the target object. While some of these cases are arguably annotation errors, in many cases it is not possible to distinguish which object is being indicated by the bounding box, as in the *bed/sleeping bag* case in Fig. 5 (image L). Referential uncertainty of this kind is a roadblock for the use of L&V resources to study naming variation. Note that pointing gestures in natural communication are as referentially uncertain as bounding boxes, if not more; however, typically those gestures are grounded in a specific discourse context, which helps reduced uncertainty. In future work, we plan to filter out these cases.

5.1. Naming variation and agreement

We analyze the response sets obtained per object, that is, the set of names and their frequency (number of annotators entering a particular name). Our analysis of naming variation shows that, on the one hand, we have a fair bit of consistency in the names chosen for objects, and, on the other, also consistent variation. Figure 6 shows the cumulative histograms for type counts, i.e. how many objects have at least n names, with different frequency thresholds t. Without any frequency thresholding t=1, that is, allowing names entered by only one annotator, the proportion of instances that have a single name annotated is very small, below 10%, and there is a long tail of datapoints with many names, up to 19. With a reasonable (based on data inspection; names entered by one annotator only have the most noise, which is to be expected) threshold of t=2, a bit over 20% of the objects have one name, almost 50% up to 3 names, and 100% up to 8 names. This threshold is the one illustrated in Fig. 5, which includes names that have at least frequency 2. The average number of

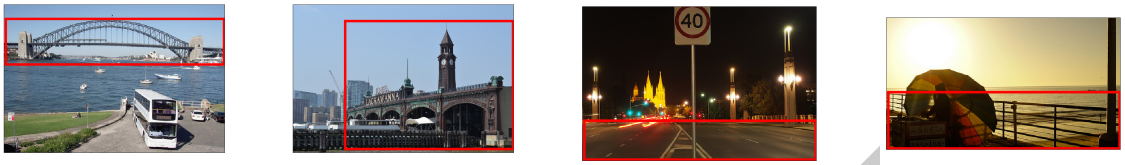


A: sandwich (34)

B: sandwich (15), basket (6), food (5), burger (2), hamburger (2), meal (2)

C: food (10), sandwich (8), toast (5), french toast (4), dessert (2), breakfast (2)

D: hotdog (14), food (7), bun (4), sandwich (3), bread (2)

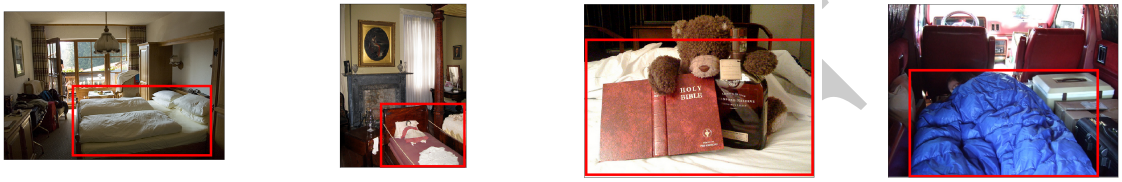


E: bridge (35)

F: bridge (20), building (11)

G: street (16), road (15), bridge (3)

H: pier (6), railing (5), dock (5), bridge (5), fence (4), rail (3), boardwalk (3)



I: bed (36)

J: bed (16), bench (6), crib (5)

K: bed (17), book (6), table (4), toy (3), bible (2), doll (2)

L: bed (12), sleeping bag (9), blanket (7), bed sheet (5)

Figure 5: Examples for VisualGenome images labeled *sandwich*, *bridge*, and *bed* (first, second, last row, respectively) with higher to lower agreement in ManyNames.

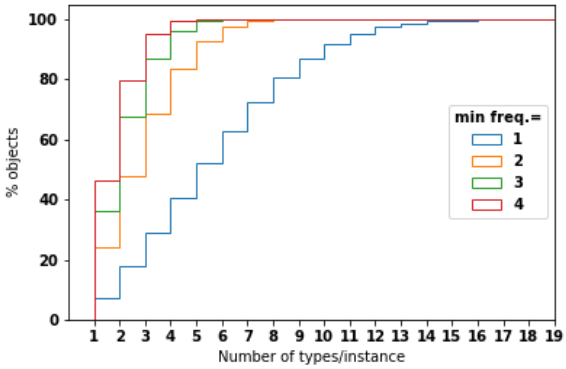


Figure 6: Cumulative histograms for number of types in ManyNames, with frequency thresholds

names with this threshold is 2.9, and the most frequent name accounts on average for 75% of the responses for a given object (Table 4, and see below). Thus, in our data objects tend to have a preferred name, as expected from work in Psychology (Rosch et al., 1976; Jolicoeur, 1984), and at the same time there is variation.

To further assess agreement on the object names, we check the following measures, computed with $t = 2$, with results in Table 4:

- **N**: the average number of types in the response set of ManyNames.

domain	N	%top (std)	H (std)	t=VG	%VG
all	2.9	75.2 (21.9)	0.9 (0.7)	72.8	62.8
people	4.3	59.0 (20.4)	1.5 (0.7)	49.8	36.3
clothing	3.2	70.1 (18.5)	1.1 (0.6)	70.2	57.4
home	3.1	72.6 (20.7)	1.0 (0.7)	78.5	64.1
buildings	3.0	74.7 (20.7)	1.0 (0.7)	72.6	61.6
food	2.9	76.4 (20.7)	0.9 (0.7)	62.9	55.2
vehicles	2.4	76.6 (19.8)	0.8 (0.6)	71.1	63.9
animal_plants	1.5	94.5 (12.1)	0.2 (0.4)	93.8	91.0

Table 4: Agreement in object naming, with a frequency threshold of 2.

- **% top**: the average relative frequency of the most frequent response (shown in percent).
- H: the H agreement measure from (Snodgrass and Vanderwart, 1980), where 0 is perfect agreement: $H = \sum_{i=1}^{k} p_i \, log_2 \frac{1}{p_i}$, where k denotes the number of name types and p_i is the proportion of name type i in the responses.
- **t=VG**: the percentage of items where the top response in ManyNames is the VG name.
- **% VG**: the average relative frequency of the VG name in the response set.

Apart from the trends mentioned above, it is remarkable that only in 73% of the cases the most frequent response

coincides with the VG name, and the VG name accounts for 63% of the responses on average. Our dataset can be expected to yield a more robust estimate for so-called entry-point names (Jolicoeur, 1984), that is, the name that most naturally comes to mind for a given object. The H measures indicate a fair amount of agreement, a bit lower than in picture norming studies on artificial idealized images (e.g. Snodgrass and Vanderwart (1980) report an average H of 0.55), which can be expected when using real images.

If we check agreement by domain, two domains stand out: the ANIMALS_PLANTS domain, which is often discussed in the literature and where we find almost perfect agreement ($H = 0.2$), and the PEOPLE domain with a particularly low agreement. Across domains, however, we find a large standard deviation for both %top and H, of around 20% for all domains except PEOPLE. This indicates that agreement varies quite a bit across instances, with factors that cannot be attributed to domain only. The qualitative examples in Figure 5 illustrate this, showing visual instances with very high or low agreement. These suggest that instances which are more prototypical of a category trigger higher agreement, although further research is necessary to examine the relevant factors. The following section will examine other sources of variation in object naming.

5.2. Sources of variation

Previous work on object naming has assumed that variation is mostly along a taxonomic axis, and in particular hierarchical (see Section 2.). This parameter does not seem to explain the variation in ManyNames. Table 5 shows the distribution of the lexical relations between ManyNames responses and the original VG annotation, estimated from WordNet. To obtain these data, we exploited the synset annotation in the VG names, and added automatic linking for the additional ManyNames names, with a simple first-sense heuristic.[4] As shown in the table, in the vast majority of cases, no hierarchical relation between the name and the synset can be retrieved from WordNet. Even factoring in the noise introduced by referential uncertainty, it is clear that a good portion of our data cannot be explained by variation in the level of abstraction of the chosen name. Among the names that do have a taxonomic relation to the synset, hypernyms are the most frequent, meaning that our annotators often went for a more general name than the VG annotators. In a qualitative analysis, we found the following types of variation in the data, illustrated with examples in Figure 5. **Cross-classification**: a substantial group are names conceptualizing alternative aspects of the same object (e.g. *toast/dessert*, image C). **Conceptual disagreement**: as we did not filter objects for prototypicality, our data mirrors a certain amount of disagreement between speakers as to what an object is (*bed/bench*, images J). **Metonymy**: we find examples reminiscent of metonymy discussed in the linguistic literature (Pustejovsky, 1991) where logically related parts of an object stand in as its name (*burger/basket*, image B). **Issues with WordNet**: due to WordNet's fine-grained hierarchy, it is difficult to retrieve certain loose synonyms or hypernyms (*robe/dress*, image not shown).

[4] To detect hypernyms, we use the hypernym closure of the synset with a depth of 10; the other relations are straightforward. The coverage of WordNet for our name data is satisfactory (90% of the name types, accounting for 97% of the tokens).

relation	% types	% tokens	*ex: jacket*
word-not-covered	10.6	2.6	*outdoor vest*
synonym	1.1	1.1	*hoodie*
hyponym	2.2	3.8	*parka*
co-hyponym	3.1	5.9	*raincoat*
hypernym	10.5	27.7	*clothing*
rel-not-covered	72.2	58.3	*sweatshirt*

Table 5: Lexical relations of naming variants in ManyNames to annotated VG synset, averaged over synsets, with examples of variants for *jacket*.

6. Conclusion

The question of how people choose names for objects presented visually is relevant for Language and Vision, Computational Linguistics, Computer Vision, Cognitive Science, and Linguistics. We have surveyed datasets that can be useful to address this question, and proposed a new dataset, ManyNames, that affords new possibilities both for analysis and modeling of object naming.

For Computer Vision and L&V, our data highlights the fact that bounding boxes are often ambiguous, which can affect model performance on object categorization and naming. Crucially, evaluations in these tasks assume that object identification is possible based on the bounding box; beyond showing that this is not always the case, our data can be used to assess whether model mistakes are plausible (similar to those of humans, as in the *toy/book/bed* case), or really off.

Moreover, standard evaluations assume that object names (or categories) are unique. The ability to distinguish incorrect object names from good alternatives is essential for visual object understanding. Our data provides a first step towards enabling model evaluation on naming variants of an instance, checking, e.g., to what extent the top N predicted names are valid alternatives (*dog, animal, pet*) or not (*dog, hat, grass*). However, to fully enable this sort of analysis, a further annotation step is needed, to account for the referential uncertainty of bounding boxes and annotation noise. We plan to take this step in future work, which will also enable more robust conclusions with respect to naming variation.

With respect to naming variation, our current data supports the prediction in theoretical research on object naming that there will often be a preferred (entry-level) name for a given visually presented object. It tentatively suggests that (a) there is also consistent variation in naming, with an average of almost three elicited names per instance; (b) much of this variation cannot be explained by adopting a hierarchical view, which has been dominant in the psycholinguistic and computational literature; (c) there is high variability in agreement across instances within the same domain. The latter suggests that there are specific visual characteristics of either the object itself or the visual context in which it appears that trigger variation. With prototypical, idealized pictures of the sort used in traditional studies (see Figure 1), this observation would not be possible.

We hope that ManyNames triggers more empirical research on object naming, a topic that has been understudied in both computational and theoretical approaches to language.

7. Bibliographical References

Bernardi, R., Cakici, R., Elliott, D., Erdem, A., Erdem, E., Ikizler-Cinbis, N., Keller, F., Muscat, A., and Plank, B. (2016). Automatic description generation from images: A survey of models, datasets, and evaluation measures. *J. Artif. Int. Res.*, 55(1):409–442, January.

Brown, R. (1958). How shall a thing be called? *Psychological review*, 65(1):14.

Das, A., Kottur, S., Gupta, K., Singh, A., Yadav, D., Moura, J. M., Parikh, D., and Batra, D. (2017). Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2.

De Vries, H., Strub, F., Chandar, S., Pietquin, O., Larochelle, H., and Courville, A. (2017). Guesswhat?! visual object discovery through multi-modal dialogue. In *Proc. of CVPR*.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

Devlin, J., Cheng, H., Fang, H., Gupta, S., Deng, L., He, X., Zweig, G., and Mitchell, M. (2015). Language models for image captioning: The quirks and what works. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 100–105, Beijing, China, July. Association for Computational Linguistics.

Fang, H., Gupta, S., Iandola, F., Srivastava, R., Deng, L., Dollar, P., Gao, J., He, X., Mitchell, M., Platt, J., Zitnick, L., and Zweig, G. (2015). From captions to visual concepts and back. In *Proceedings of CVPR*, Boston, MA, USA, June. IEEE.

Fellbaum, C. (1998). *WordNet*. Wiley Online Library.

Graf, C., Degen, J., Hawkins, R. X., and Goodman, N. D. (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.

Humphreys, G. W., Riddoch, M. J., and Quinlan, P. T. (1988). Cascade processes in picture identification. *Cognitive neuropsychology*, 5(1):67–104.

Jolicoeur, P. (1984). Pictures and names: Making the connection. *Cognitive psychology*, 16:243–275.

Kazemzadeh, S., Ordonez, V., Matten, M., and Berg, T. L. (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.

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

Krause, J., Johnson, J., Krishna, R., and Fei-Fei, L. (2017). A hierarchical approach for generating descriptive image paragraphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 317–325.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., Bernstein, M., and Fei-Fei, L. (2016). Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations.

Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Duerig, T., and Ferrari, V. (2018). The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv:1811.00982*.

Lazaridou, A., Dinu, G., and Baroni, M. (2015). Hubness and pollution: Delving into cross-space mapping for zero-shot learning. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 270–280, Beijing, China, July. Association for Computational Linguistics.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., DollÃ¡r, P., and Zitnick, C. (2014). Microsoft coco: Common objects in context. In *Computer Vision - ECCV 2014*, volume 8693, pages 740–755. Springer International Publishing.

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., and McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of ACL: System Demonstrations*, pages 55–60.

Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A. L., and Murphy, K. (2015). Generation and comprehension of unambiguous object descriptions. *CoRR*, abs/1511.02283.

McRae, K., Cree, G. S., Seidenberg, M. S., and McNorgan, C. (2005). Semantic Feature Production Norms for a Large Set of Living and Nonliving Things. *Behavior Research Methods*, 37(4):547–559.

Murphy, G. (2004). *The big book of concepts*. MIT press.

Olson, D. R. (1970). Language and thought: Aspects of a cognitive theory of semantics. *Psychological review*, 77(4):257.

Ordonez, V., Liu, W., Deng, J., Choi, Y., Berg, A. C., and Berg, T. L. (2016). Learning to Name Objects. *Commun. ACM*, 59(3):108–115, February.

Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., and Lazebnik, S. (2015). Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models. *CoRR*, abs/1505.04870.

Pont-Tuset, J., Gygli, M., and Ferrari, V. (2019). Natural vocabulary emerges from free-form annotations. *arXiv preprint arXiv:1906.01542*.

Pustejovsky, J. (1991). The generative lexicon. *Computational linguistics*, 17(4):409–441.

Rohde, H., Seyfarth, S., Clark, B., Jäger, G., and Kaufmann, S. (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*, pages 107–116.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M.,

and Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive psychology*, 8(3):382–439.

Ross, B. H. and Murphy, G. L. (1999). Food for thought: Cross-classification and category organization in a complex real-world domain. *Cognitive psychology*, 38(4):495–553.

Rossion, B. and Pourtois, G. (2004). Revisiting snodgrass and vanderwart's object pictorial set: The role of surface detail in basic-level object recognition. *Perception*, 33(2):217–236.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252.

Shafto, P., Kemp, C., Mansinghka, V., and Tenenbaum, J. B. (2011). A probabilistic model of cross-categorization. *Cognition*, 120(1):1 – 25.

Snodgrass, J. G. and Vanderwart, M. (1980). A standardized set of 260 pictures: norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of experimental psychology: Human learning and memory*, 6(2):174.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *CVPR 2015*, Boston, MA, USA, June.

Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Yu, L., Poirson, P., Yang, S., Berg, A. C., and Berg, T. L., (2016). *Modeling Context in Referring Expressions*, pages 69–85. Springer International Publishing.

Zhao, H., Puig, X., Zhou, B., Fidler, S., and Torralba, A. (2017). Open vocabulary scene parsing. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2002–2010.