

Do Objects in Real-World Images Have a Canonical Name?

Anonymous EMNLP-IJCNLP submission

Abstract

Research in language & vision commonly acknowledges the fact that speakers can interpret and describe visual scenes in many different ways. For instance, frameworks in image captioning typically elicit (during collection) or train and test on (during modeling) multiple valid alternatives. Individual visual objects and their names, however, are often collected and modeled in simple annotation set-ups where annotators or systems identify objects via bounding boxes and label them with a single “correct” name. A prime example here is VisualGenome that features complex and varied annotations on the level of scenes and regions, but provides canonicalized objects linked to a single name. In this paper, we put the assumption that objects in complex visual scenes bear a canonical name to an empirical test. Based on pre-annotated object bounding boxes in VisualGenome, we elicit multiple names from 36 subjects per object and investigate the extent to which there is variation in the names chosen by different people for the same object. Our analysis reveals that non-canonicalized object naming data shows a lot of interesting linguistic variation as speakers name objects on different levels of specificity (cow-animal) or verbalize different aspect of the same object (bowl-salad). At the same time, we find that object naming prompted via bounding boxes is subject to a certain amount of noise as speakers have problems re-identifying the object that was annotated by the original bounding box. We investigate whether a state-of-the-art model of object labeling implicitly encodes similar variation in object naming and discuss implications for research in language & vision.

1 Introduction

Much of the recent progress in Language & Vision research goes back to large-scale image annotation efforts where human annotators have (i)

segmented images into regions and objects and (ii) labelled these regions with words or phrases. This annotation-based approach to data collection is typically useful for obtaining massive resources that provide enough training data for learning e.g. object labels, region descriptions, etc.

Expressions describing or referring to objects in visual scenes typically include object names: e.g., *cheesecake* or *dessert* in Figure ?? . Determining these object names is a core aspect of virtually every language & vision task, ranging from e.g. referring expression generation to visual dialogue (?). We investigate the extent to which there is variation in the names chosen by different people for the same object, and its implications for research in language & vision.

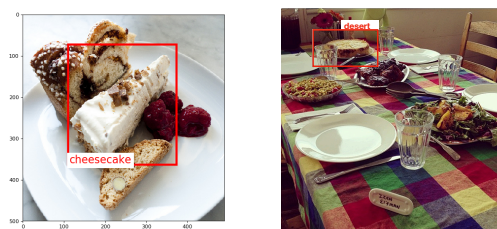


Figure 1: Two objects of the same type of cake, with different names in VisualGenome

Our paper puts together two strands of research that have mostly been pursued independently to date. On the one hand, state-of-the-art computer vision systems are able to accurately classify images into thousands of different categories (e.g. ?), where the task is often to predict the name for a given object. //g: *Is this true? Imagenet task asks for synsets, which can be taken to be categories...To refine//* However, they mostly adopt very simple assumptions with respect to the underlying lexicon, which is implemented as a simple, flat labeling scheme: A standard object recognition system would be trained to classify the left

object in Figure 1 as *cheesecake*, the right one as *dessert*, and using *dessert* for the left picture would be considered incorrect. On the other hand, research on object naming in Cognitive Science has shown that people choose different names depending on the circumstances, with factors such as context or the prototypicality of the object with respect to the category playing a role (?). *//g: This research also argues that there is high agreement in how people name objects; to do: make coherent.* However, this research typically uses stylized drawings are used, and is focused on taxonomic relations (*sparrow-bird*). *//sz: It is thus unclear how findings from these stylized settings generalize to tasks in language & vision like referring expression generation, where naming is a core aspect. Therefore, in contrast to traditional naming norm studies in Cognitive Science we study object naming in realistic scenes where objects are situated in a natural context! (This comes with additional challenges, like potential object occlusion, background/foreground confusion etc.)*

In our study, we collect large-scale object naming data via crowdsourcing. Like object naming studies in Cognitive Science, we collect multiple names per object (concretely, 36); like most work on language & vision, we use natural images *//sz: (showing objects in complex visual contexts, surrounded by other objects, not ImageNet-like images)* on a large scale, annotating objects in 25K images from the Visual Genome dataset. We analyze the agreement in object naming across subjects, and the sources of variation. We find that: *//g: To be put in paragraph form*

- there is quite a high level of agreement in the task, with the relative frequency of the most common name being 70% on average. This is in accordance with previous results in Cognitive Science (?);
- the level of agreement in object naming is much higher in certain domains than in others; as it happens, the domains that have been traditionally used in object naming research (e.g. animals) seem to display the highest amount of agreement in our data set;
- most of the variation in our dataset comes from alternative names that do not stand in a taxonomic relation, suggesting that the previous work in Cognitive Science is missing much of the empirical ground.

our datasets contains a lot of variability for names coming from different parts of the taxonomy (*dessert* vs. *cake*, *bottle* vs. *wine*)

Moreover, we analyze whether current models implicitly encode the variation in naming, by doing XXX. We find YYY.

2 Related Work

Cognition: Concepts and categorization Seminal work on concepts by Rosch and colleagues suggests that object names typically exhibit a preferred level of specificity, which ? called the **entry-level**. This typically corresponds to an intermediate level of specificity, i.e., **basic level** (e.g. *bird*, *car*) (?), as opposed to more generic (i.e., **super-level**; e.g., *animal*, *vehicle*) or specific categories (i.e., **sub-level**; e.g., *sparrow*, *convertible*). However, less prototypical members of basic-level categories tend to be instead identified with sub-level categories (e.g., a PENGUIN is typically called a *penguin* and not a *bird*) (?). 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 are flexible in their choice of the level of specificity. Scenarios where multiple objects (of the same category) are present induce a pressure for generating names which uniquely identify the target (?), such that sub-level names can be systematically elicited in these cases (??).

Vision: Object Recognition State-of-the-art computer vision systems are able to classify images into thousands of different categories (e.g. ?). These object recognition systems are now widely used in vision & language research. Nevertheless, the way the treat object recognition is conceptually very simple (if not to say, naive): standard object classification schemes are inherently “flat”, and treat object labels as mutually exclusive (?), ignoring all kinds of linguistic relations between these labels and ignoring the fact that an object can easily be an instance of several categories. *//cs: I would make this statement stronger and argue that object recognition is merely a labeling of objects with human interpretable symbols, and that a system would probably fail if it had to decide whether an object labeled as, e.g. fig may also be labeled as food.* *//g: ok*

Vision & language: Naming and Referring ? have studied the problem of deriving appropriate

object names, or so-called entry-level categories, from the output of an object recognizer. Their approach focusses on linking abstract object categories in ImageNet to actual words via translation procedures that e.g. involve corpus frequencies. ? learn a model of object naming on a corpus of referring expressions paired with objects in real-world images, but focus on combining visual and distributional information and on zero-shot learning. Object naming is also an important task for referring expression generation, though most research in this area has focussed on content and attribute selection (????). *//g: Should we put this as motivation also in the intro?//*

3 Data collection

We take data from VisualGenome (?), which aims to provide a full set of descriptions of the scenes which images depict in order to spur complete scene understanding. It contains a dense region-based labeling of 108k images with textual expression of the attributes and references of objects, their relationships as well as question answer pairs, all linked to WordNet synsets (? , see below).

3.1 Sampling of Instances (Images/Objects)

We aimed at collecting a relatively large amount of naturalistic images (*instances*) that depict common objects. *//sz: say more clearly again that we have complex images with many objects and collect names for particular objects in these images?//* We start from the concepts of McRae et al.'s feature norms (REF), which are common objects of different categories (e.g., ANIMALS, FURNITURE) and, as such, have a high overlap with standard datasets of object norming studies (REFS). We added the PERSON category because it is very frequent category in VisualGenome.

(As appropriate: 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).)

Collection nodes We defined a set of *collection nodes* which we would then use to collect our object instances from VG.

We based the definition of our set of nodes on the WN (REF) synsets of the McRae concepts (e.g., dog, duck, goose, gull), the nominal WordNet hierarchy, and the frequency distribution of the VG object names' synsets.¹

¹TODO: need to be clear from the general description of

First, we selected a set of collection node candidates—synsets which match (e.g., *dog*, *duck*, *goose*, *gull*) or subsume (e.g., *mammal*, *bird*) the McRae synsets². From these candidates we kept those as collection nodes which had a high frequency of VG object instances of different names. For example, VG instances subsumed by McRae's *dog* were named *beagle*, *greyhound*, *puppy*, *bull-dog*, etc., while McRae's *duck*, *goose*, or *gull* did not have name variants in VG, so we kept *dog* and *bird* as collection nodes.

Collection of instance candidates Goal of above procedure was the collection of instances of selected object classes—our nodes— whose VG names correspond to or subsume (are hypernyms of) a McRae concept, and whose object names differ, that is, of which we can expect that people possess different names for them (e.g., *duck*, *goose*, *gull* for *bird*). The collection of such instances using the nodes was then straightforward: We retrieved all VG images depicting an object whose name matches or is subsumed by one of the collection nodes. We did not consider objects with names in plural form, with parts-of-speech other than nouns³, or that were multi-word expressions/phrases (e.g., *pink bird*). We further only considered objects whose bounding box⁴ had an area of 20 – 90% of the whole image area.

Sampling of instances Finally, from this set of instances we sampled our final dataset of 31,093 instances. Sampling proceeded in dependence on the overall size of the individual collection seeds: up to 800 objects per seed: all instances, but at most 500, are collected; more than 800 objects per seed: all instances, but at most 1,000, are collected. **double-check**

//cs: END @ GBT//

Table ?? gives an overview of the collection nodes, XXX, XXX, grouped into 7 domains. **(Report only dataset after round0, with a note in caption/footnote referring to the checkpoint pruning.)** *//g: I would put only a table with the final dataset, not the intermediate one (round0). That is, include a table with the 25K imgs (the 30K imgs were just our initial pool).//*

Number of images/objects: 25,596

Number of object names: 450

VG that the frequ. of instances labeled with the synset of the object name is meant.

²Specific synset IDs, e.g., dog.n.01, are omitted for read-

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)
truck (465)	cheese (110)	cow (324)	couch (366)	overpass (23)	guy (330)	t-shirt (62)
boat (450)	donut (78)	bird (295)	chair (365)	grill (22)	child (230)	tie (51)
jet (106)	salad (70)	sheep (216)	carpet (307)	garage (18)	batter (110)	blazer (43)
aircraft (85)	sauce (68)	bull (48)	bowl (219)	hotel (16)	kid (85)	hood (26)
van (76)	apple (33)	flower (40)	curtain (182)	castle (14)	skateboarder (80)	cap (20)

Table 1: Overview of our dataset: Top-10 VG names for each domain (number of instances in parentheses).

double-check

animals_plants	vehicles	home	clothing	buildings	food
ungulate.n.01 (2037)	aircraft.n.01 (1208)	furnishing.n.02 (5355)	shirt.n.01 (968)	house.n.01 (364)	dish.n.02 (812)
horse.n.01 (833)	train.n.01 (957)	vessel.n.03 (525)	overgarment.n.01 (786)	bridge.n.01 (297)	baked_goods.n.01 (770)
feline.n.01 (763)	car.n.01 (727)	kitchen_utensil.n.01 (132)	dress.n.01 (199)	shelter.n.01 (169)	foodstuff.n.02 (280)
dog.n.01 (688)	motorcycle.n.01 (564)	crockery.n.01 (92)	headress.n.01 (135)	restaurant.n.01 (58)	vegetable.n.01 (448)
bird.n.01 (389)	truck.n.01 (559)	cutlery.n.02 (82)	neckwear.n.01 (65)	outbuilding.n.01 (31)	edible_fruit.n.01 (42)
flower.n.01 (44)	boat.n.01 (499)	tool.n.01 (72)	robe.n.01 (27)	hotel.n.01 (19)	beverage.n.01 (23)
rodent.n.01 (27)	ship.n.01 (38)	lamp.n.01 (34)	glove.n.02 (7)	housing.n.01 (17)	
insect.n.01 (12)			footwear.n.01 (5)	place_of_worship.n.01 (12)	
fish.n.01 (11)					

Table 2: Overview of our dataset: Collection nodes for each domain (number of instances in parentheses). **double-check**

Number of collection nodes (synsets): 52

plural form, most frequent names is of same domain as VG name (gives 25, 596, i.e., discard 5, 497 instances)

3.2 Procedure

describe the crowdsourcing set-up and the task
 TODO: Footnote: we ran pilot experiments to design our experiment and instructions.

- workers could only participate in one round, such as to avoid workers annotating an instance more than once.

Collection Method

- instructions;
Appendix A gives the instructions as they were presented to the annotators.
- each round: HIT of 10 instances, collect 9 annotations for each HIT
- round 0 (with opt-outs) → pruning → rounds 1-3 (no opt-outs)
pruning: Based on given opt-outs: keep images with no OCCLUSION, at most BBOX is ambiguous twice, at most 17% of names in

ability.

³(REF to tagger)

⁴TODO: need to be clear from the general description of VG what is meant.

Overall XX participants, each annotated between XX and XX instances. *//g: Put mean or median instead of min-max//*

3.3 Data

give an overview of the collected data

4 Analysis

4.1 Agreement

We compute the following agreement measures:

- % top:** for each object, we calculate the relative frequency of the most common name, and then average over all objects
- SN:** for each object, we calculate the Snodgrass agreement measure, and then average

over all objects *//g: Note: changing SD to SN cause SD is typically standard deviation//*

- **=VG**: the proportion of objects where the most frequent name coincides with the name annotated in VisualGenome

Table 4 shows that, overall, our annotators achieve a fair amount of agreement in the object naming choices. The domain where annotators agree most is the animal domain, which, interestingly, happens to be the domain that has been mostly discussed in the object naming literature. *//sz: ... much more to say//*

Why is naming more flexible in certain domains than in others? *//g: Hypothesis: expectation: little variation - hypernymy at most, more variation j-i more affordances j-i more varied relationships.//*

4.2 Lexical relations

In this section, we take a closer look at the lexical variation we observe in our data set. We analyze the data points where participants attributed different names to the same object and extract a set of pairwise **naming variants**. These naming variants correspond to pairs of words that can be used interchangeably to name certain objects. For each object, we extract the set of naming variants $s = \{(w_{top}, w_2), (w_{top}, w_3), (w_{top}, w_4), \dots\}$ where w_{top} is the most frequent name annotated for the object and $w_2 \dots w_n$ constitute the less frequent alternatives of w_{top} . The **type frequency** of a naming variant (w_{top}, w_x) corresponds to the number of objects where this variant occurs. The **token frequency** of (w_{top}, w_x) corresponds the count of all annotations where w_x has been used instead of w_{top} . In Table 6, we show the the naming variants with the highest raw token frequency for each domain.

The naming variants can be grouped according to their lexical relation, as follows:

- **synonymy**: e.g. aircraft vs. airplane
- **hyponymy**: e.g. man vs. person
- **co-hyponymy**: e.g. swan vs. goose
- **no relation**: e.g. desk vs. apple

Research on object naming following the idea of entry-level categories has, essentially, exclusively looked at names that stand in a hierarchical relation (i.e. hyponymy/hypernymy).

We use WordNet to extract lexical relations between the naming variants in our data set. Unfortunately, this means that we have to exclude a certain portion of the data as either (i) one of the name is not covered in WordNet, (ii) we cannot find a lexical relation between the two names (see below). Also, we had to be relatively permissive with respect to the definition of hyponymy/co-hyponymy. For instance, to analyze *giraffe* as a hyponym of *animal* we have to look at the closure of the hyponyms of *animal* with a depth of 8 (in WordNet). *//sz: should we call this co-hyponymy or co-hierarchical relation?//*

//sz: include Table that reports counts of the naming variants, coverage in WordNet etc.// //g: I think it'd be best to put the out-of-wordnet info in the Lexical relations table – this way we have everything in one place.//

Table 5 shows the distribution of lexical relations for those naming variants that we were able to analyze with WordNet. Both in terms of their types and token frequency, the naming variants that instantiate a (loose) co-hyponymy relation are by far the most frequent. *//sz: discuss in more detail, discuss: to what extent is this an artefact of WordNet?//* This is really interesting: most research on object naming, to date, has focussed on hyponymy/hypernymy, i.e. variation that relates to hierarchical relations between object names. Our data suggests that co-hierarchical variation is really important too.

4.3 Crossclassified

Table

Interesting name pairs:

- storefront - store: strictly speaking it's part-whole, but how can one distinguish between the two?
- field - grass: same (reverse); how to distinguish?
- dog - pet (different conceptualizations; classified as "overlapping")
- airplane - flight, plane - flight (classified as "adjacent" although it's an event relation...)

5 Modeling

6 Conclusions

A Instructions for AMT Experiment

domain	non-canonicalized			annotated vs. non-canonicalized		
	% top	SD	N _{types}	Max=VG	% VG (Micro)	% VG (Macro)
people	51.88	2.13	8.60	49.77	32.25	32.25
clothing	63.93	1.58	6.44	70.16	52.58	52.63
home	66.35	1.50	6.31	78.48	58.82	58.92
buildings	66.87	1.55	6.89	72.60	55.46	55.59
food	71.27	1.30	5.49	62.89	52.11	52.25
vehicles	71.97	1.13	4.66	71.11	60.18	60.20
animals_plants	91.26	0.44	2.68	93.84	88.01	88.01
all	69.69	1.34	5.70	72.77	58.65	58.43

Table 3: Agreement in non-canonicalized naming data, and between annotated (original VisualGenome) and non-canonicalized names for different object domains

domain	% top	SD	Max=VG	% VG	max synset	%	SD	Max=VG	% VG	min synset
people	51.88	2.13	49.77	32.25	professional.n.01	60.67	2.02	20.00	11.67	athlete.n.01
clothing	63.93	1.58	70.16	52.58	neckwear.n.01	79.32	0.91	76.92	65.45	footwear.n.01
home	66.35	1.50	78.48	58.82	tool.n.01	86.44	0.73	94.44	82.34	crockery.n.01
buildings	66.87	1.55	72.60	55.46	bridge.n.01	75.48	1.21	86.87	69.58	place_of_worship.n.01
all	69.69	1.34	72.77	58.65	None	None	None	None	None	None
food	71.27	1.30	62.89	52.11	edible_fruit.n.01	79.67	0.89	78.57	69.78	vegetable.n.01
vehicles	71.97	1.13	71.11	60.18	train.n.01	92.74	0.42	99.16	92.14	aircraft.n.01
animals_plants	91.26	0.44	93.84	88.01	feline.n.01	94.78	0.29	98.95	94.15	fish.n.01

Table 4: Agreement in object names for objects of different domains, if applicable, synsets with maximal and minimal agreement (top %) are shown

domain	crossclassified		co-hyponymy		hypernymy		not-covered		synonymy	
	typ	tok	typ	tok	typ	tok	typ	tok	typ	tok
people	0.725	0.618	0.019	0.032	0.051	0.293	0.200	0.038	0.005	0.019
clothing	0.709	0.661	0.020	0.046	0.045	0.195	0.219	0.085	0.008	0.012
animals_plants	0.679	0.461	0.032	0.102	0.111	0.365	0.167	0.058	0.011	0.014
food	0.590	0.433	0.031	0.033	0.104	0.432	0.267	0.089	0.009	0.013
vehicles	0.635	0.329	0.026	0.083	0.059	0.239	0.271	0.089	0.008	0.260
home	0.672	0.644	0.026	0.072	0.040	0.130	0.254	0.097	0.009	0.057
buildings	0.780	0.725	0.026	0.038	0.045	0.156	0.138	0.058	0.011	0.022
all	0.731	0.574	0.033	0.057	0.076	0.256	0.147	0.050	0.013	0.064

Table 5: Lexical relations between naming variants according to WordNet, for the set of name pairs where both words can be found in WordNet and stand in a *//sz: should we produce this table for the different domains?// //g: yes, please. Maybe do rows domains, columns lexical relations (synonymy, hyponymy, co-hyponymy, other, not in wordnet), with subcolumns for types and tokens? And do percentages over rows – for each domain, how many of the variants we find fall into each of the classes. This way we’ll be able to see differences across domains.//*

Task:
Please name the object in the red box with the **first name that comes to mind**.

- Make sure to identify the correct object: **the single object that the box marks**. It is the one that the **box fits tightly around**.
- If you cannot name the object, click one of the options below the textbox.

Make sure to avoid the **mistakes** exemplified below:





	<p>Object Name: <input type="text" value="man"/></p> <p>If you cannot name the object, please specify the reason:</p> <p>Image quality:</p> <p><input type="radio"/> Object occluded / not recognizable</p> <p><input type="radio"/> Bounding box is unclear</p> <p><input type="radio"/> Other: <input type="text" value="Please specify the reasons"/></p>	<p>BAD---the box is marking the jacket, not the man</p> <p>Good options (examples): jacket, coat, shirt</p>
	<p>Object Name: <input type="text" value="man"/></p> <p>If you cannot name the object, please specify the reason:</p> <p>Image quality:</p> <p><input type="radio"/> Object occluded / not recognizable</p> <p><input type="radio"/> Bounding box is unclear</p> <p><input type="radio"/> Other: <input type="text" value="Please specify the reasons"/></p>	<p>BAD---the box is marking the car, not the motorcycle</p> <p>Good options (examples): car, suv, automobile</p>
	<p>Object Name: <input type="text" value="Enter an object name"/></p> <p>If you cannot name the object, please specify the reason:</p> <p>Image quality:</p> <p><input checked="" type="radio"/> Object occluded / not recognizable</p> <p><input type="radio"/> Bounding box is unclear</p> <p><input type="radio"/> Other: <input type="text" value="Please specify the reasons"/></p>	<p>OK, too (since the car is occluded)</p>
	<p>Object Name: <input type="text" value="hard hat"/></p> <p>If you cannot name the object, please specify the reason:</p> <p>Image quality:</p> <p><input type="radio"/> Object occluded / not recognizable</p> <p><input type="radio"/> Bounding box is unclear</p> <p><input type="radio"/> Other: <input type="text" value="Please specify the reasons"/></p>	<p>BAD---this is not a name, only hat is.</p>


Figure 2: Instructions for AMT annotators for round 0.

YouNameIt Instructions (Click to collapse)

Task: Please name the object in the red box with the **first name that comes to mind**.

- Make sure to identify the correct object: **the single object that the box marks**. It is the one that the **box fits tightly around**.
- For more advice please see the FAQ below.


Make sure to avoid the **mistakes** exemplified below:



Object Name:

BAD---the box is marking the jacket, not the man


Good options (examples):
jacket, coat, shirt, clothing, ...



Object Name:

BAD---the box is marking the car, not the motorcycle

Good options (examples):
car, suv, vehicle, automobile, ...



Object Name:

BAD---this is not a name, only hat is.

Good options (examples):
hat, helmet, headgear...

FAQ:

+ If there is a person riding a skateboard or riding a surfboard, would you prefer the answer to be *man* or *surfer*?

There is no unique "correct" or "best" answer---we ask you to enter the first name that comes to mind for the marked object. For example, if you first think of *surfer*, then please enter this name. If, however, *man* comes to mind first, then enter *man*.

+ For food, for example, a photo of a pizza would be *pizza* and not *food*?

That's up to you---there is no unique "correct" or "best" answer. Instead, please enter the name which first occurred to you for the identified object; if you first thought *food*, then *food*, if first *pizza*, then *pizza*.

+ I was wondering what you wanted us to do with images and boundary boxes around women, men, children that are not particularly doing anything? It looks as though you do not want us to name it *man* or *woman*...

Any name is ok as long as it identifies the right object; *man*, *woman*, *child*, etc. are all valid names for objects. It is important, though, to make sure that you have identified the correct object for which we ask you to give a name. As the instructions say, it is the object that the bounding box is tightly around.

+ In some cases, a group of objects together could be described with a singular noun (i.e. *bundle*, *salad*, *sandwich*, etc.).

If you can describe the objects with a singular noun, i.e. they together form some kind of object, such as sandwich or salad, then this is completely fine. It is important though that this "group of objects" is to be named, that is, the bounding box is tightly around them as a whole (for instance, pay attention that it is not only marking the filling of a sandwich).

+ When there is a picture of a person and the whole person is in the red box and they are doing a sport is it better that I say *player*, *snowboarder*, *skier*, etc." instead of *man*, *woman*, *girl*, *boy*, etc?

There is no "correct answer", so you are just asked to enter the first name that comes to mind for the marked object. Referring directly to your example, any of the options, *player* OR *snowboarder* OR *man* OR *woman* etc., are good.

If your question is not clarified by the FAQ, feel free to contact us (Carina.Silberer"at"upf.edu).

Figure 3: Instructions for AMT annotators for rounds 1 to 3.

category	most frequent naming variants
people	woman – person (3594), man – person (3546), boy – child (3243), woman – girl (2328), girl – child (1985), woman – tennis player (1277), man – player (1273), man – boy (1214), skateboarder – skater (1194), man – t-shirt (1143)
food	pizza – food (1883), sandwich – food (1123), hotdog – food (540), pizza – cheese (457), pizza – plate (430), salad – food (402), sandwich – burger (398), hotdog – sandwich (351), sandwich – bread (318), cake – food (286)
home	couch – sofa (4090), desk – table (3448), carpet – floor (1697), bench – chair (1401), desk – keyboard (1380), counter – table (1201), table – desk (1135), counter – countertop (1101), table – counter (906), rug – carpet (895)
buildings	house – building (1160), building – house (511), bridge – train (326), bridge – overpass (235), house – window (161), house – home (123), tent – canopy (120), building – castle (101), bridge – building (98), bridge – pole (85)
vehicles	airplane – plane (11194), plane – airplane (3829), motorcycle – bike (2624), airplane – jet (1319), boat – ship (1301), truck – car (1095), car – vehicle (874), motorcycle – wheel (861), truck – vehicle (718), truck – wheel (716)
clothing	shirt – t-shirt (2914), jacket – coat (2396), jacket – shirt (1552), jacket – suit (1168), suit – jacket (1029), shirt – jacket (813), shirt – tie (723), shirt – man (487), shirt – dress (462), shirt – sweater (450)
animals_plants	cow – bull (515), sheep – goat (486), cow – animal (445), giraffe – animal (380), bird – parrot (349), sheep – animal (294), sheep – lamb (282), horse – animal (269), cat – animal (237), bird – seagull (231)

Table 6: Most frequent naming variants for each category

same object (217; 31%)			not same object (483; 69%)		
co-hyponyms	104	horse-donkey, truck-jeep, castle-church	adjacent	167	building-clock, cat-pole, bed-chair
overlapping	69	skater-boy, cat-toy, horse-statue	supports.1	122	couch-pillow, bowl-fruit, person-dress
hypernym.1	18	hat-beanie, jacket-tuxedo, building-hut	part.2	102	plane-wings, pizza-cheese
hypernym.2	16	donut-pastry, train-vehicle, van-automobile	supports.2	64	shirt-child, donut-plate, carpet-floor
other	8	sandwich-meal, plane-plain, umbrella-tent	part.1	20	wheel-moped, sink-counter, bread-sub
			other	10	plane-flight, horse-wood

Table 7: Relations between words or objects when the variants refer to the same object (left) or not (right).