Object naming is context dependent - A case study on testing linguistic hypotheses on real-world image corpora

Anonymous ECCV submission

Paper ID ***

Abstract. The abstract should summarize the contents of the paper. LNCS guidelines indicate it should be at least 70 and at most 150 words. It should be set in 9-point font size and should be inset 1.0 cm from the right and left margins. ...

Keywords: We would like to encourage you to list your keywords within the abstract section

1 Introduction

Understanding and modeling the way humans converse about their environment using natural language has been a long-standing goal of research in various fields related to artificial intelligence and linguistics. But, for a long time, computational research into language grounding, reference and situated interaction had to rely on "block-worlds" or very controlled, experimental environments [1–5], leading to models and corresponding findings at a rather restricted scale. This situation has radically changed with recent developments in the language & vision community. Pushed especially by developments in computer vision, there has been an explosion of interest in language & vision tasks, ranging from image captioning [6–10], referring expression resolution and generation [11–14], to multi-modal summarization or visual dialogue [15, 16].

Despite this dramatic progress in the field, however, a lot of research in language & vision still centers around systems and applications for very specific tasks and data sets. Most of the influential papers have investigated which modeling architectures need to be adopted to re-generate human descriptions or utterances in a particular data set, whereas studies aimed at obtaining linguistic generalizations from existing data collections and models are relatively rare.

In this paper, we take a look at a straightforward task in language & vision that so far has found surprisingly little attention in the community: object naming, i.e. determining the nominal that speakers will use to refer to an object in a visual scene. With state-of-the-art computer vision systems being able to classify images into thousands of different categories [17], it might even seem like this task is already solved. However, theoretically, the lexical choice involved in naming is a non-trivial one [18]. Real-world objects are members of many categories and speakers can typically choose between different, more or less specific

names when referring to a particular visual entity. For instance, the entity surrounded by the green box in Figure 1 is an instance of the categories *female child, child, female, person, organism*, etc. and can be referred to with names such as, e.g., *girl, kid, cutie, daughter, person, human*. Most research on computer vision and object recognition has somehow worked around the fact that objects can be instances of multiple, hierarchically embedded categories. State-of-the-art recognition models are mostly trained as classifiers on a flat of 1000 categories in ImageNet, reducing the ImageNet ontology to an inherently flat annotation scheme.

The goal of this paper is to provide a systematic discussion of the object naming task and to show that existing, large-scale resources do not meet the requirements for a theoretically informed model of this very basic phenomenon in language & vision.

2 Naming Objects (in Context)

TODO@Laura

2.1 Background

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") and can be seen as a subtask of generating a referring expression. The lexical choice involved in naming is a non-trivial one [18]. Indeed, since an object is simultaneously a member of multiple categories (e.g., a young beagle is at once a dog, a beagle, an animal, a puppy etc.), all the various names that lexicalize these constitute a valid alternative: in fact, their denotation includes the target object. The study of object naming focuses on those factors that lead to the choice of a particular name in communicative situations.

Lexical alternatives in naming differ in their **level of specificity** (e.g., *dog* is less specific than *beagle*)[19]: the categories they denote can indeed be organized in a hierarchical fashion according to class inclusion relations (e.g., a beagle is a dog, a dog is an animal etc.) [20]. Such a conceptual structure gives rise to a taxonomy, or ontology. It was observed that each type of object exhibits a preferred level of specificity which it is more naturally named at, called the **entry-level**. This typically corresponds to an intermediate level of specificity, i.e., **basic level** (e.g, *bird*, *car*) [21], 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*) [22]. This out-of-context preference towards a certain taxonomic level is often referred to as **lexical availability**.

Contextual factors also affect object naming. Scenarios where multiple objects are available induce a pressure for generating names which uniquely identify

the target, thus excluding the competing alternatives (or *distractors*)[23]. For example, in presence of more than one dog, the name *dog* is ambiguous in terms of the specific one it is intended to refer to. In these cases, a sub-level category (e.g., *rottweiler*, *beagle*) is more informative: it provides more specific information, though it is possibly more costly to produce (it may be a non-default, infrequent or long alternative). It was shown that in use speakers tend to choose a name given a trade-off between its **cost**, on one side, and its contextual **informativeness**, on the other; for example, they may still opt for an ambiguous expression if less costly for them to produce it [24] [25]. Other contextual factors affecting lexical choice include the **perceptual salience** of the object, such as its size or location [26].

2.2 Previous Work

As mentioned above, there is hardly any work on language & vision that has studied object naming explicitily, on large-scale data sets. [27] investigate 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 various translation procedures. Their performance is surprisingly low: around 35% precision. We need to discuss this paper in more detail.

3 Requirements

In order to study object naming as prompted by real-world images, not only do we need to collect names naturally generated by speakers, but also to quantify those factors whose interaction with naming we want to analyze. As we saw, phenomena that impacts on object naming are both perceptual and linguistic in nature: since we here focus on perception as vision, we distinguish between those that pertain language, vision or both.

//la: Note that I swapped the order of contextual informativeness and lexical availability. What does R1 refer to? the requirement of sub-level info?//

— **Visual-Linguistic factors**

R1 **Lexical availability**:

In object naming, there is a preference for each object type towards a certain level of specificity. To detect this, we need to have access to a taxonomy, reflecting hierarchical relations among the various categories, and map both names and target objects to their location in such a structure. For instance, if we want to check whether a convertible is more often referred to with a basic-level category, e.g., *car*, or a more specific one, e.g., *convertible*, we need to know that *car* lexicalizes a category which is super-ordinate to that lexicalized by *convertible*. Moreover, in the first place, we need to know which categories our target object belongs to (e.g., convertible, a car, a vehicle etc.) and hence which names could

apply to it. Note that for due to class inclusion relations, it is sufficient to know the most specific category in the taxonomy, i.e., the sub-level, to deduce all the other others (e.g., convertibles $\subset$ cars $\subset$ vehicles ...). For this purpose, one can use existing lexical resources as a taxonomy: for example, in WordNet [28] words denoting the same category are grouped and linked to others by taxonomic relations. The repertory of categories, and hence candidate names, for a target object can be queried by taking the sub-level category, i.e., the *leaf category* in the taxonomy.

R2 **Contextual informativeness**:
The same object may be named differently depending on the distractors it is combined with: a particular name may not be sufficiently specific to identify its referent if applicable to more than one object. In order to analyze this, we are required to know the repertory of categories that each of the potential referents is an instance of: for instance, we need to know that there are multiple cars, to judge whether *car* is ambiguous. As before, we also need to map names to a taxonomy, in order to check which level of specificity is chosen. We then require similar resources as in R1, but extending information about the sub-level category to each object in the scene.

— **Linguistic factors**:

R3 **Cost**:
The cost of generation of a nominal may impact on its likelihood to be chosen in naming. Previous work [25] operationalized such cost in terms of length and frequency. These can readily be coded as functions of respectively the number of characters and the relative frequency (estimated from a text corpus) of a name. *//la: Actually, they also include typicality in the cost but I would leave it out as we kind of put that phenomenon more under the umbrella of lexical availability. And in general, I am not sure what to say about that, actually ... Do we want typicality ratings?//* However, reasoning about the cost of lexical alternatives requires having access to such alternatives, hence names applicable to the object. As before, these can be obtained by mapping the object to the taxonomy and derive the set of categories it is a member of.

— **Visual factors**

R4 **Perceptual salience**:
The prominence of an object in a scene may affect naming. This can be estimated geometrically on the basis of the relative position and size of an object in the image.

To summarize, to carry out an analyses of object naming in real-world images datasets we require:

- Names naturally generated by speakers to refer to a target object in an image
- Sub-level category of each object in the image (including the target object)
- Set of categories mapped to their lexical realizations and organized according to taxonomic relations (e.g., WordNet)
- Coordinates of the image region showing the target object.

Fig. 1. INCLUDE MSCOCO examples

Resource	Purpose	#Cat	#Ref/ object	#Refs	Constraints on language	Constraints on images	Coordinates
RefCOCO	REG,REC	80	3	50k	-	≥2 objs of same cat	✓
RefCOCO+	REG,REC	80	3	50k	+	≥2 objs of same cat	✓
Flickr30k Entities	Phrase localization	?			–		✓
VisualGenome	Caption generation Scene understanding Phrase localization	80k					✓

Table 1. Overview of V&L benchmarks for taks related to reference. Cat: Categories; Ref: Reference; REG and REC: Referring expression generation and comprehension, respectively.

4 Resources: What Do We Have?

TODO@Carina

4.1 Linguistic Resources, Computer Vision Resources

WordNet WordNet [28] blabla.

Methods: Object Detectors, Image Classifiers *//cs: Do we need that? I would say no.//*

ImageNet and ILSVRC

MS COCO [29] MS COCO is an object recognition dataset of images of natural scenes. As such it provides region-level object annotations for 91 common object categories (e.g., DOG, PIZZA, CHAIR). Multiple datasets for vision & language tasks have been built on top of COCO, such as the referring expressions datasets RefCOCO and RefCOCO+ which we will discuss below, and *COCO Captions* [30]. The latter provides five captions for each of 300k images, spanning 80 of

Resource	Provides (Undirectly)			
	R1: Most specific	R2: Entry-level	R3: unconstraint	WordNet ID
RefCOCO	✗	✗	✗	✗
RefCOCO+	✗	✗	✗	✗
Flickr30k Entities	✗	(✓)	✓	✗
VisualGenome	✗	✓	✓	✓

Table 2. BLABLA overview of resources and their shortcomings wrt object naming. REG and REC: Referring expression generation and comprehension, respectively.

the COCO categories. However, the COCO object annotations are not linked to the captions.

4.2 V & L Resources

//cs: [If Table 1 is kept (?):]
Table 1 gives an overview of the resources we discuss.//

RefCOCO and RefCOCO+ [13]
Both datasets are extensions of ReferIt [11], a large-scale collection of referring expressions (RE) for natural objects in real-world images, and are built on top of the MS COCO image collection [29].

The REs were collected via crowdsourcing in a two-player game that was designed to obtain REs which uniquely refer to the target objects in an image. Specifically, a director and a matcher are presented with an image, and the director produces a RE for an outlined target in the image. The matcher must click on the object he thinks the RE refers to. If the matcher's prediction is correct, the RE is considered valid. For more details on the datasets see [13].

Finally, ReferIt is based on the SAIAPR image collection [31] and captures not only REs for objects ("things") but also for scene categories (e.g.,*grass, road*) (see also [32, 33]). The task of scene detection is beyond the focus of our study regarding object naming, we therefore will not discuss this dataset further.

Flickr30k Entities
The Flickr30k Entities dataset [34][1] augments Flickr30k, a dataset of 30k images and five sentence-level captions for each of the images, with region-level annotations. 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. The dataset was designed to advance image description generation and phrase localization in particular (e.g., [35–37]).

VisualGenome
VisualGenome [38] 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

[1] Available at web.engr.illinois.edu/ bplumme2/Flickr30kEntities

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 [28, see below].

4.3 Discussion of Shortcomings

We discuss the shortcomings of the presented V&L datasets with respect to their use for computational linguistic studies of reference in general, and in particular in how far they satisfy the Requirements R1–R3 for the study of object naming which we put forward in Section 3.

The REs for both, *RefCOCO+* and *RefCOCO* (henceforth, RefCOCO /+), were collected under the constraints that (i) all images contain at least two objects of the same category (80 COCO categories), which prompts the players to avoid the mere object category as RE, and (ii) in RefCOCO+ the players must not use location words, urging them to refer to, e.g., the appearance of objects. While these constraints ensured that the datasets are interesting from the computer vision perspective, they fall short in containing phenomena that are intriguing from the view of language research, and are in contradiction to Requirement R3: First, how the choice of a RE for an object interacts with the categories of its distractors can only partially be observed in the data[2] due to (i). And second, the study of how people *naturally* refer to objects requires that speakers are not constraint in their choice of REs, which (ii) does not fulfill.

Furthermore, (iii), not all objects in the images were annotated with REs, may it due to the frequency constraint (i), or due to the object not being part of the 80 COCO categories. For this reason and the fact that RefCOCO /+ the 80 COCO categories tend to be entry-level categories, it does thus not fulfill requirement R1. *//cs: TODO: EX / ANALYSIS//*

The dataset also falls short in fully meeting requirement R2 due to property (iii)—it only covers 80 object categories—and constraints (i,ii) to some extent—we may not be able to reliably infer the entry-level from the data *//cs: TODO: Example or remove last part//*.

By design, *Flickr30k Entities* can be used to study the way people refer to individual entities in an image in dependence on the situation the speakers describe. The annotators were not constraint by interfering external factors in their choice of entity mentions. *//cs: double-check?//* We can therefore gather naturally produced object names from the data for objects which are mentioned in the captions and may be able to infer their entry-level categories (Requirement R2). *//cs: ADD EXAMPLE IMAGE//*

Flickr30k Entities has similar issues in satisfying R1 as RefCOCO /+. Object categories tend to be even less specific than those of COCO (e.g., PEOPLE, ANIMALS, BODYPARTS, CLOTHING), or are abstract (OTHER, SCENE).

[2] For example, the preference of the entry-level category over a sufficiently unique more generic category can only be observed in images in which the target and the distractor objects are of different more generic categories.

[EN$_{39}$ Two people] in [EN$_{45}$ the photo] are playing
[EN$_{40}$ the guitar] and [EN$_{41}$ the other] is poking at [EN$_{0}$ him] .
[EN$_{42}$ A man] in [EN$_{43}$ green] holds [EN$_{40}$ a guitar] while
[EN$_{41}$ the other man] observes [EN$_{43}$ his shirt] .
[EN$_{41}$ A man] is fixing [EN$_{43}$ the guitar players costume] .
[EN$_{41}$ a guy] stitching up [EN$_{43}$ another man 's coat] .
[EN$_{39}$ the two boys] playing [EN$_{40}$ guitar]

Fig. 2. Example of Flickr30k Entities .

Annotation types	
Regions	dalmatian's head; dalmatian dog wating at a dinner table; the dog is gray
Attributes	dog is gray; dalmation is eating
Relationships	dog at table; dalmation IN chair; heeler eating treat

Fig. 3. Example of VisualGenome .

Note that Flickr30k Entities is less suited for referring expression generation and interpretation*//cs: use the term grounding, interpretation or understanding to refer to the process of linking language to an image region?//*, though, in that the mentions in isolation of their linguistic context may not uniquely identify the referred object. For example, the two men shown in the image in Figure 2 are referred to by *a man* or *a guy*.

VisualGenome From a linguistic perspective, object categories (e.g., FAWN, INDUSTRIAL PARK, OPENED BAG, CARPET ORIENTAL) are defined pragmatically, and the annotations are rather unstructured. For example, as illustrated in Figure 3, the categorization of expressions into one of the three annotation types (region, attributes and relationships), is based on the verbs (*is* denotes attributes) and prepositions (*at* or *in* denotes relationships) *//cs: double-check w/ paper//* Note also the spelling mistakes (e.g., *dalmation*, *wating*).

As far as object naming is concerned, requirement R1 is not met. The annotators were free in their choice of a name for the objects, hence, the latter is usually the entry-level in many annotations. For the same reason, on the other hand, and because of the dense image labeling and its exhaustive annotations, VisualGenome satisfies R2—we can infer the entry-level name of each object in the image. XX (Figure 3)

Table 2 summarizes the shortcomings of the resources with respect to our requirements.

	#synset(*obj*)	#*obj* for which ∗(synset(*obj*), ILSVRC)		
		in hypernym		hyponym
RefCOCO (REs)	2,580	112	204	36
RefCOCO (categories)	78	15	30	1 (?)
VisualGenome				
Flickr30k Entities	6,139	238	338	95

Table 3. Coverage of the unique object names (obtained from the categories or from the REs) in the datasets by WordNet (first column) and by the 1,000 synsets of the ILSVRC image classification challenge (columns 2–4). #synsets(*obj*): no. names for which synsets could be retrieved from WordNet;

Solution to R1: Object detectors? An alternative would be to apply object detectors or image classifiers trained to predict the most specific category of the full inventory of objects which the dataset covers. However, pre-trained models only exist for a subset of the datasets' objects. *//cs: TODO: ADD CMP WITH ILSVRC//* For the training of a model using, e.g., ImageNet [39], on the other hand, the set of most-specific categories covered by the data needs to be provided or collected from humans.

Solution to R2: Captions? Regarding RefCOCO /+, additional REs could be collected from the COCO Captions dataset. Since annotators could naturally refer to depicted objects in their descriptions (i.e., choose the entry-level as the most preferred reference whenever possible), we may infer the entry-level of the objects from them through maximum likelihood estimation. In contrast to Flickr30k Entities , though, the data does not contain region-phrase associations, such that natural language phrases first needed to be aligned with the image regions they refer to. This task of *language grounding*, which has been an active research topic in vision & language (e.g., [40, 41, 35]), is beyond the focus of our object naming study.

//cs: XXXXX//

4.4 Analysis of Data wrt Our Requirements(?)

Pre-processing We parse referring expressions and captions with the Stanford Dependency Parser. We extract heads/object names as follows: TODO.

Level of Specificity Variability of reference level in existing data sets for language & vision? Are resources appropriate for defining reference level?

WordNet? We hypothesize that the distance of a name's synset to the root node (entity) relates to its specificity. We estimate this distance as the minimal path length of all synsets of a word to the root node.

Table 4 shows the estimated levels of specificity for object names in the RefCoco data set. We observe distances to the root between 2 and 17, meaning that there is a much more fine-grained distinction of levels than the three-way classification.

Unfortunately, the levels of specificity predicted by WordNet do not seem to reflect linguistic intuitions, here are some problematic examples from Table 4:

– elephant (10) is more specific than panda (14)? horse is less specific than elephant (10)?

specificity	rel.freq.	top 5 names
-1	0.071697	NONE,brocolli,zeb,broc,girafe
2	0.003898	thing,things
3	0.001182	object,group,set,substance,objects
4	0.140633	man,person,piece,head,part
5	0.100739	player,glass,baby,front,corner
6	0.208590	woman,girl,kid,boy,bowl
7	0.238708	guy,right,chair,lady,bear
8	0.110613	horse,bus,cow,pizza,batter
9	0.097390	shirt,car,bike,donut,catcher
10	0.048368	elephant,couch,truck,vase,suitcase
11	0.008828	motorcycle,clock,mom,dad,scissors
12	0.002822	oven,airplane,suv,taxi,refrigerator
13	0.005253	laptop,fridge,canoe,orioles,pigeon
14	0.000414	panda,freezer,penguin,rooster,rhino
15	0.030870	zebra,giraffe,zebras,giraffes,deer
16	0.000083	bison,mooses,orang,elks,sambar
17	0.000143	ox,cattle,gnu,mustang,orca

Table 4. Levels of specificity for naming choices in RefCOCO: for each level, relative frequency and 5 most frequent names are shown

No. of images in ImageNet?

5 Proposal

TODO@Sina (?)

vision people have focussed on learning very specific categories, and have not cared about learning actual concepts/names,

language & vision people have focussed on learning what speakers say in certain data sets, we know that speakers will use object names at a medium level of specificity most of the time, so it is difficult to learn the meaning of

specific concepts from existing resources and to implement a model that can decide which level in the taxonomy is appropriate in context — we need a more systematic approach for collecting object naming data, controlling the level of specificity (e.g. prompting speakers to come up with the most specific word they can think of)... this data might also give us insight into what basic-level and sub-level concepts are ..

yes, will try ...

References

1. Anderson, A.H., Bader, M., Bard, E.G., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., et al.: The hcrc map task corpus. Language and speech **34**(4) (1991) 351–366
2. Fernández, R., Schlangen, D.: Referring under restricted interactivity conditions. In Keizer, S., Bunt, H., Paek, T., eds.: Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue, Antwerp, Belgium (September 2007) 136–139
3. Krahmer, E., Van Deemter, K.: Computational generation of referring expressions: A survey. Computational Linguistics **38**(1) (2012) 173–218
4. Takenobu, T., Ryu, I., Asuka, T., Naoko, K.: The rex corpora: A collection of multimodal corpora of referring expressions in collaborative problem solving dialogues. (2012)
5. Zarrieß, S., Hough, J., Kennington, C., Manuvinakurike, R., DeVault, D., Fernandez, R., Schlangen, D.: Pentoref: A corpus of spoken references in task-oriented dialogues. In: 10th edition of the Language Resources and Evaluation Conference. (2016)
6. Fang, H., Gupta, S., Iandola, F., Srivastava, R., Deng, L., Dollar, P., Gao, J., He, X., Mitchell, M., Platt, J., Zitnick, L., Zweig, G.: From captions to visual concepts and back. In: Proceedings of CVPR, Boston, MA, USA, IEEE (June 2015)
7. Devlin, J., Cheng, H., Fang, H., Gupta, S., Deng, L., He, X., Zweig, G., Mitchell, M.: 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), Beijing, China, Association for Computational Linguistics (July 2015) 100–105
8. Chen, X., Lawrence Zitnick, C.: Mind's eye: A recurrent visual representation for image caption generation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 2422–2431
9. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: Computer Vision and Pattern Recognition. (2015)
10. Bernardi, R., Cakici, R., Elliott, D., Erdem, A., Erdem, E., Ikizler-Cinbis, N., Keller, F., Muscat, A., Plank, B.: Automatic description generation from images: A survey of models, datasets, and evaluation measures. J. Artif. Int. Res. **55**(1) (January 2016) 409–442
11. Kazemzadeh, S., Ordonez, V., Matten, M., Berg, T.L.: ReferItGame: Referring to Objects in Photographs of Natural Scenes. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2014), Doha, Qatar (2014) 787–798
12. Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A.L., Murphy, K.: Generation and comprehension of unambiguous object descriptions. CoRR **abs/1511.02283** (2015)
13. Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T.L. In: Modeling Context in Referring Expressions. Springer International Publishing, Cham (2016) 69–85
14. Schlangen, D., Zarriess, S., Kennington, C.: Resolving references to objects in photographs using the words-as-classifiers model. In: Proceedings of the 54rd Annual Meeting of the Association for Computational Linguistics (ACL 2016). (2016)
15. Das, A., Kottur, S., Gupta, K., Singh, A., Yadav, D., Moura, J.M., Parikh, D., Batra, D.: Visual dialog. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Volume 2. (2017)

16. De Vries, H., Strub, F., Chandar, S., Pietquin, O., Larochelle, H., Courville, A.: Guesswhat?! visual object discovery through multi-modal dialogue. In: Proc. of CVPR. (2017)
17. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: CVPR 2015, Boston, MA, USA (June 2015)
18. Brown, R.: How shall a thing be called? Psychological review **65**(1) (1958) 14
19. Cruse, D.A.: The pragmatics of lexical specificity. Journal of linguistics **13**(2) (1977) 153–164
20. Murphy, G.: The big book of concepts. MIT press (2004)
21. Rosch, E., Mervis, C.B., Gray, W.D., Johnson, D.M., Boyes-Braem, P.: Basic objects in natural categories. Cognitive psychology **8**(3) (1976) 382–439
22. Jolicoeur, P.: Pictures and names: Making the connection. Cognitive psychology **16** (1984) 243–275
23. Olson, D.R.: Language and thought: Aspects of a cognitive theory of semantics. Psychological review **77**(4) (1970) 257
24. Rohde, H., Seyfarth, S., Clark, B., Jäger, G., Kaufmann, S.: Communicating with cost-based implicature: A game-theoretic approach to ambiguity. In: Proceedings of the 16th Workshop on the Semantics and Pragmatics of Dialogue. (2012) 107–116
25. Graf, C., Degen, J., Hawkins, R.X., Goodman, N.D.: Animal, dog, or dalmatian? level of abstraction in nominal referring expressions. In: Proceedings of the 38th annual conference of the Cognitive Science Society, Cognitive Science Society (2016)
26. Clark, H.H., Schreuder, R., Buttrick, S.: Common ground at the understanding of demonstrative reference. Journal of verbal learning and verbal behavior **22**(2) (1983) 245–258
27. Ordonez, V., Liu, W., Deng, J., Choi, Y., Berg, A.C., Berg, T.L.: Learning to name objects. Commun. ACM **59**(3) (February 2016) 108–115
28. Fellbaum, C.: WordNet. Wiley Online Library (1998)
29. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollr, P., Zitnick, C.: Microsoft coco: Common objects in context. In: Computer Vision ECCV 2014. Volume 8693. Springer International Publishing (2014) 740–755
30. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollr, P., Zitnick, C.L.: Microsoft COCO Captions: Data Collection and Evaluation Server. CoRR **abs/1504.00325** (2015)
31. Grubinger, M., Clough, P., Müller, H., Deselaers, T.: The IAPR TC-12 Benchmark: A New Evaluation Resource for Visual Information Systems. In: Proceedings of the International Conference on Language Resources and Evaluation (LREC 2006). (2006) 13–23
32. Hu, R., Xu, H., Rohrbach, M., Feng, J., Saenko, K., Darrell, T.: Natural language object retrieval. CoRR **abs/1511.04164** (2015)
33. Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A., Murphy, K.: Generation and Comprehension of Unambiguous Object Descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016), Las Vegas, Nevada (2016)
34. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models. CoRR **abs/1505.04870** (2015)
35. Rohrbach, A., Rohrbach, M., Hu, R., Darrell, T., Schiele, B.: Grounding of textual phrases in images by reconstruction. In: Proceedings of the European Conference on Computer Vision (ECCV 2016). (2016)

36. Plummer, B.A., Mallya, A., Cervantes, C.M., Hockenmaier, J., Lazebnik, S.: Phrase Localization and Visual Relationship Detection with Comprehensive Image-Language Cues. In: Proceedings of the International Conference on Computer Vision (ICCV 2017). (2017) 1946–1955
37. Yeh, R.A., Do, M.N., Schwing, A.G.: Unsupervised Textual Grounding: Linking Words to Image Concepts. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018). (2018)
38. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., Bernstein, M., Fei-Fei, L.: Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. (2016)
39. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: CVPR09. (2009)
40. Kong, C., Lin, D., Bansal, M., Urtasun, R., Fidler, S.: What are You Talking About? Text-to-Image Coreference. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2014). (2014)
41. Karpathy, A., Fei-Fei, L.: Deep Visual-semantic Alignments for Generating Image Descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015). (2015) 3128–3137