Multimodal CL and NLP: Combining Language and Vision for (Computational) Semantics

Carina Silberer

IMS, University of Stuttgart

CL Fall School 2024, Passau
16 - 20 September 2024

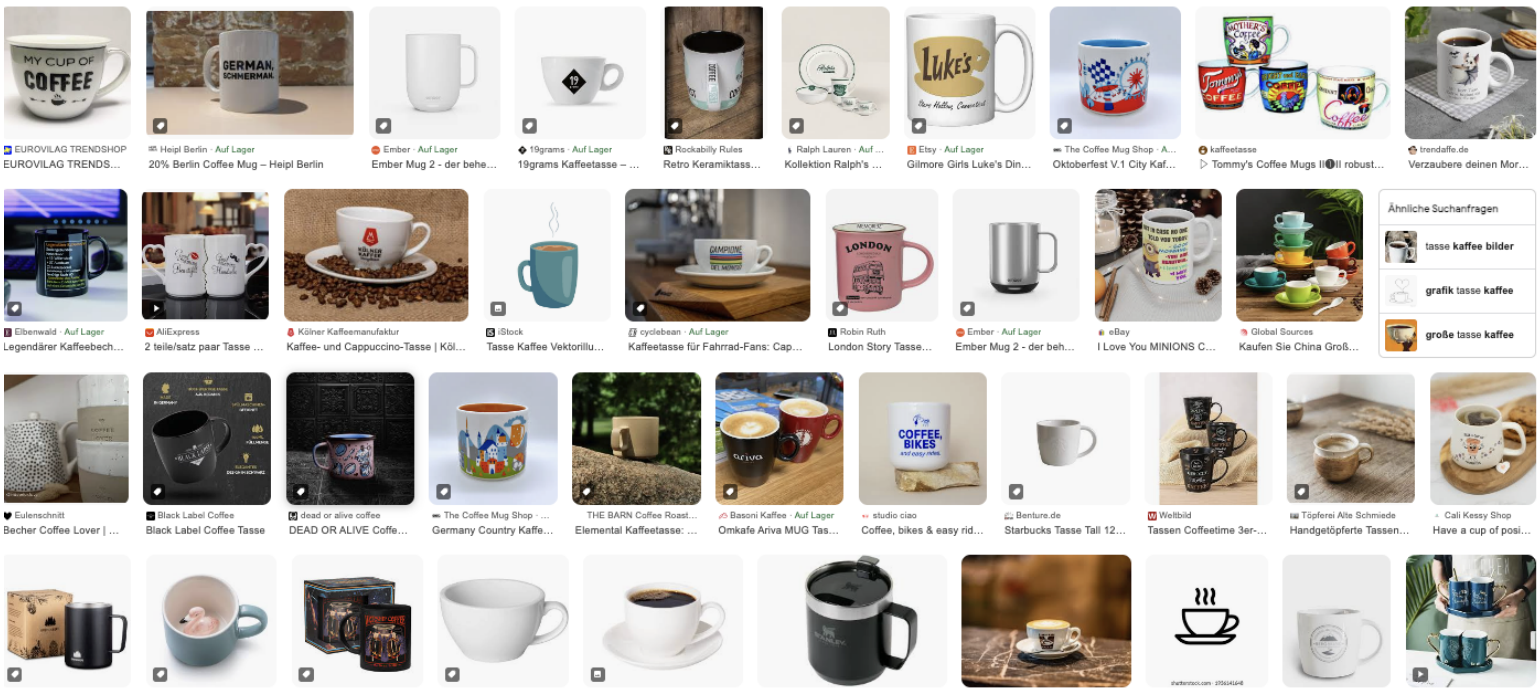
Outline

Outline

Categorisation/Object Naming/Referring Expressions

Overview

- Mapping a rich, often continuous, set of objects into a relatively small number of words (Harnad, 1990; Rosch, 1999)
- A key aspect of human language



cup, mug, coffee mug, teacup

(a small open container usually used for drinking; usually has a handle; "he put the cup back in the saucer"; "the handle of the cup was

Categorisation/Object Naming/Referring Expressions

Categorisation

- Important concept in cognitive linguistics
- Prototype model of categories (Rosch, Mervis): Categories have fuzzy boundaries; categories have central and peripheral members; members do not all share the same features
- Many-to-one
- Hierarchically related

Categorisation/Object Naming/Referring Expressions

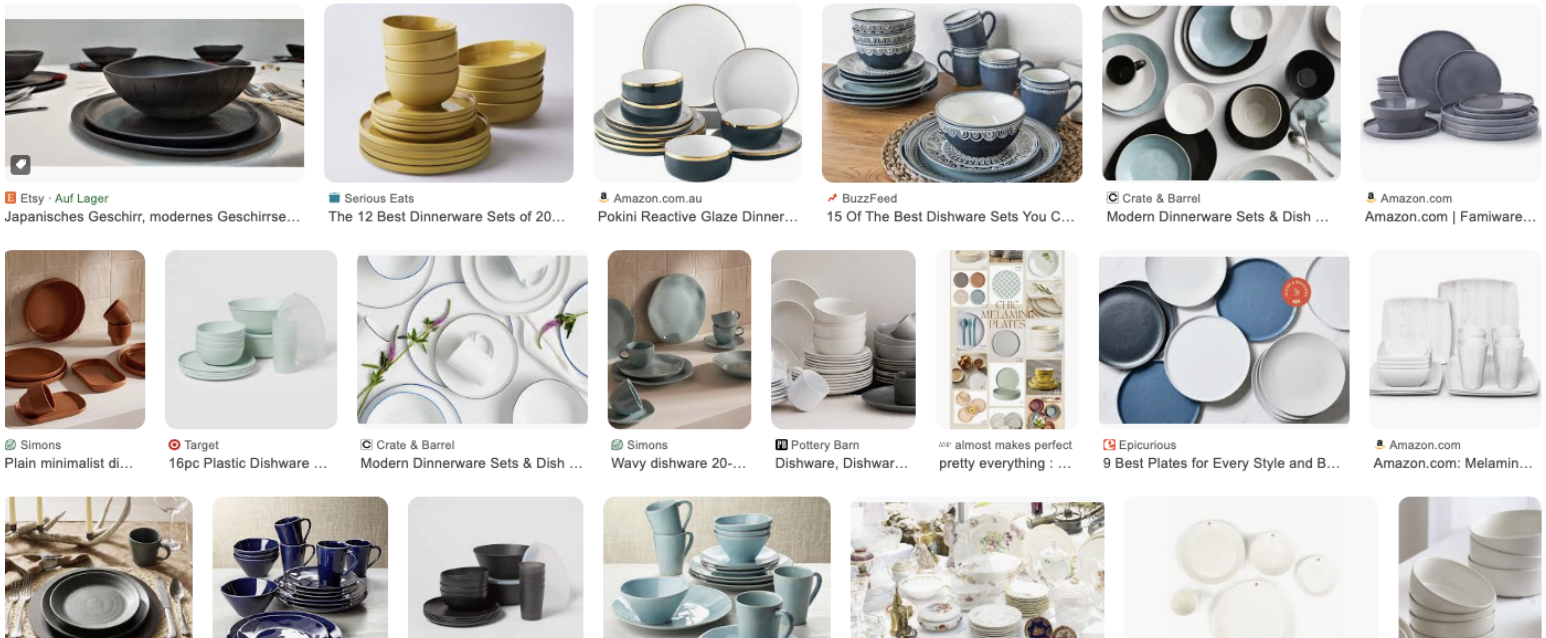
Categorisation

- Important concept in cognitive linguistics
- Prototype model of categories (Rosch, Mervis): Categories have fuzzy boundaries; categories have central and peripheral members; members do not all share the same features
- Many-to-one
- Hierarchically related

Categorisation/Object Naming/Referring Expressions

Categorisation

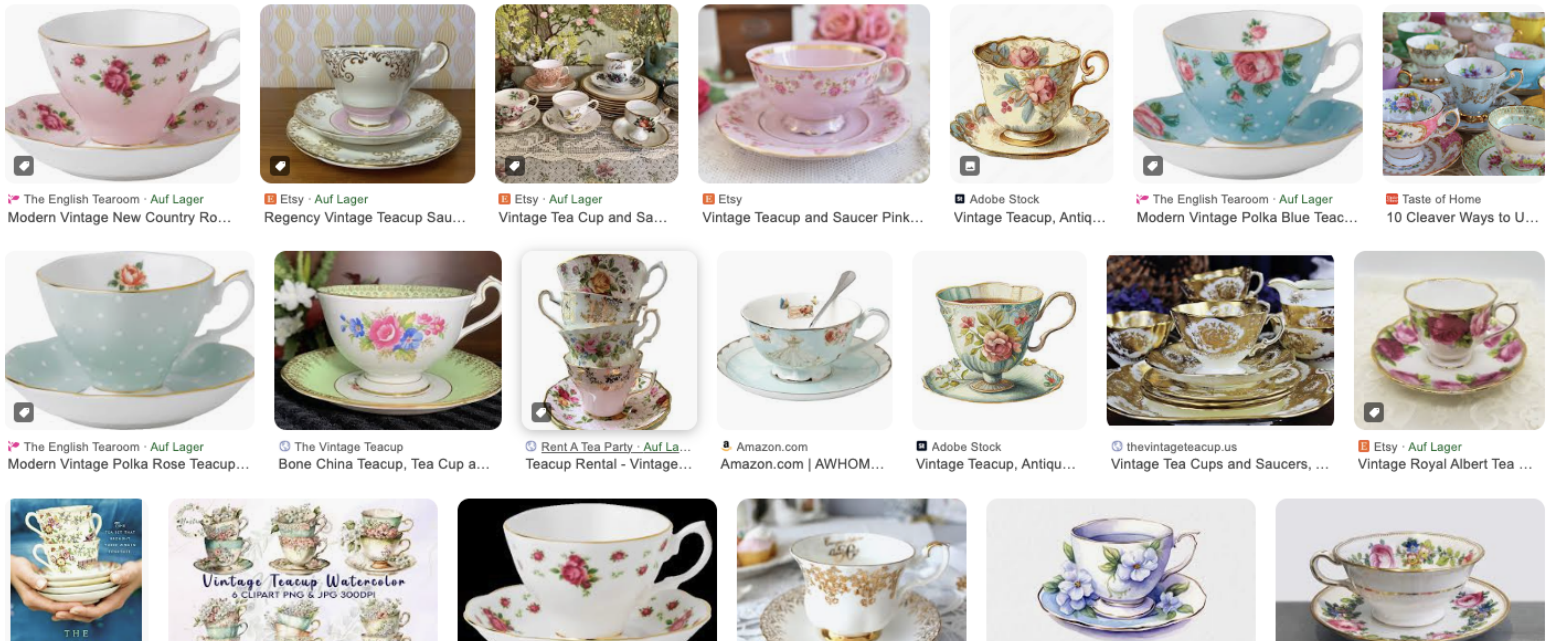
- Important concept in cognitive linguistics
- Prototype model of categories (Rosch, Mervis): Categories have fuzzy boundaries; categories have central and peripheral members; members do not all share the same features
- Many-to-one
- Hierarchically related
  supercategory (more general, hypernym)



Categorisation/Object Naming/Referring Expressions

Categorisation

- Important concept in cognitive linguistics
- Prototype model of categories (Rosch, Mervis): Categories have fuzzy boundaries; categories have central and peripheral members; members do not all share the same features
- Many-to-one
- Hierarchically related
  subcategory (more specific, hyponym)



Categorisation/Object Naming/Referring Expressions

Categorisation

- Important concept in cognitive linguistics
- Prototype model of categories (Rosch, Mervis): Categories have fuzzy boundaries; categories have central and peripheral members; members do not all share the same features
- Many-to-one
- Hierarchically related
- Works in NLP/CogSci: `ModellingSemanticCategoriesUsingConceptualNeighborhood`
- Works in CV: `https://openaccess.thecvf.com/content/CVPR2022/papers/Vaze_Generalized_Category_Discovery_CVPR_2022_paper.pdf`
- Works in VL: `https://www.nature.com/articles/s41467-020-18946-z.pdf`

Categorisation/Object Naming/Referring Expressions

Referring Expressions

- Works in NLP: Modelling Semantic Categories Using Conceptual Neighborhood bouraoui2020modelling
- Works in CV: Used for object segmentation/identification
- Works in VL:
  https://aclanthology.org/2020.acl-main.644.pdf
  https://aclanthology.org/2021.emnlp-main.516.pdf
  https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2023.1067125/full

Categorisation/Object Naming/Referring Expressions

Object Naming

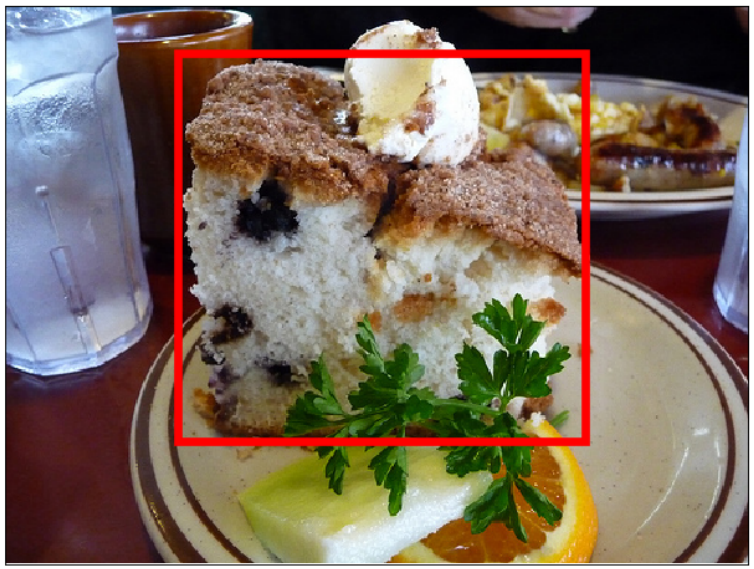
Object Naming in Language & Vision Research (L&V)

[Silberer et al., 2020a, Silberer et al., 2020b]

- Modeling how humans use language in the visual world is at the core of L&V research
- How do speakers naturally name or refer to visual objects?
- Task of object naming: assumes that objects have a canonical category or name

!?! ignores linguistic variation

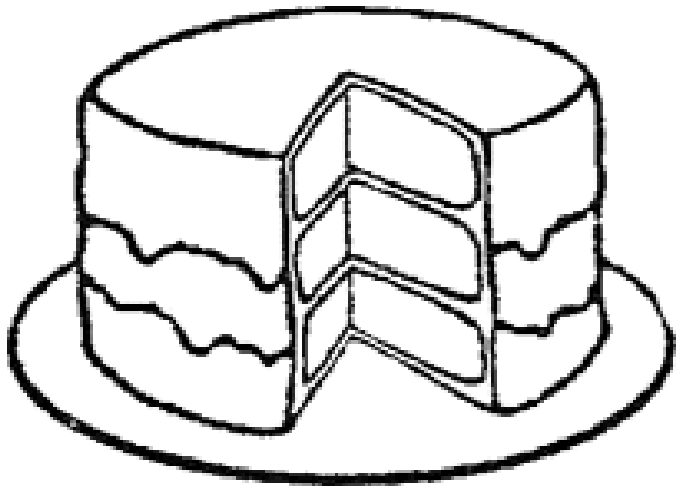
VisualGenome (VG, Krishna et al.,í6)



cake

Picture Naming Experiments in Cognitive Science

[Snodgrass and Vanderwart, 1980]



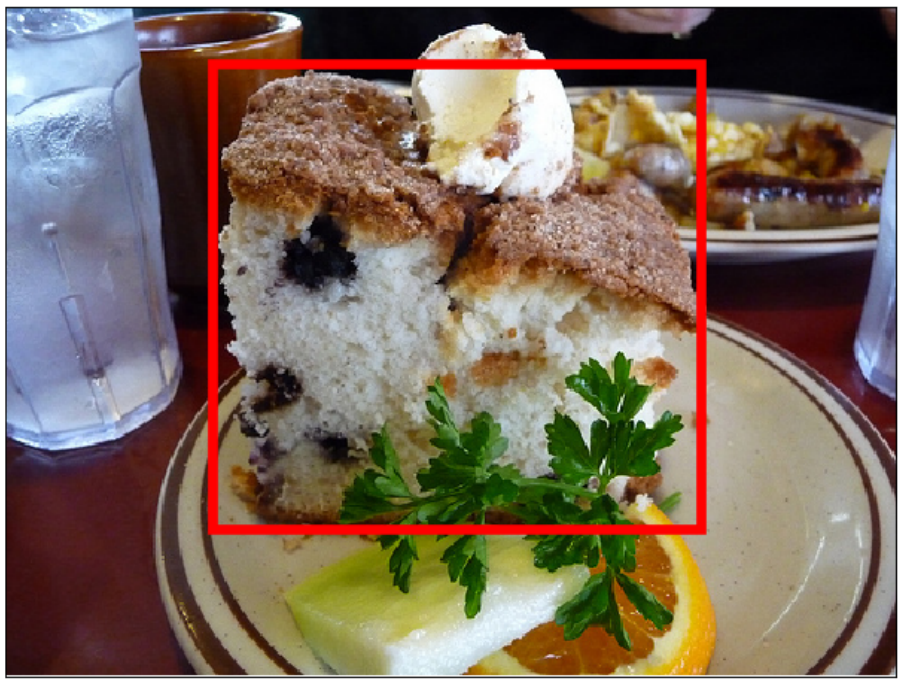
cake (83%)

Object Naming in Real-World Images

[Silberer et al., 2020a, Silberer et al., 2020b]

Phenomenon of Object Naming: Naming Variation

VisualGenome+ManyNames



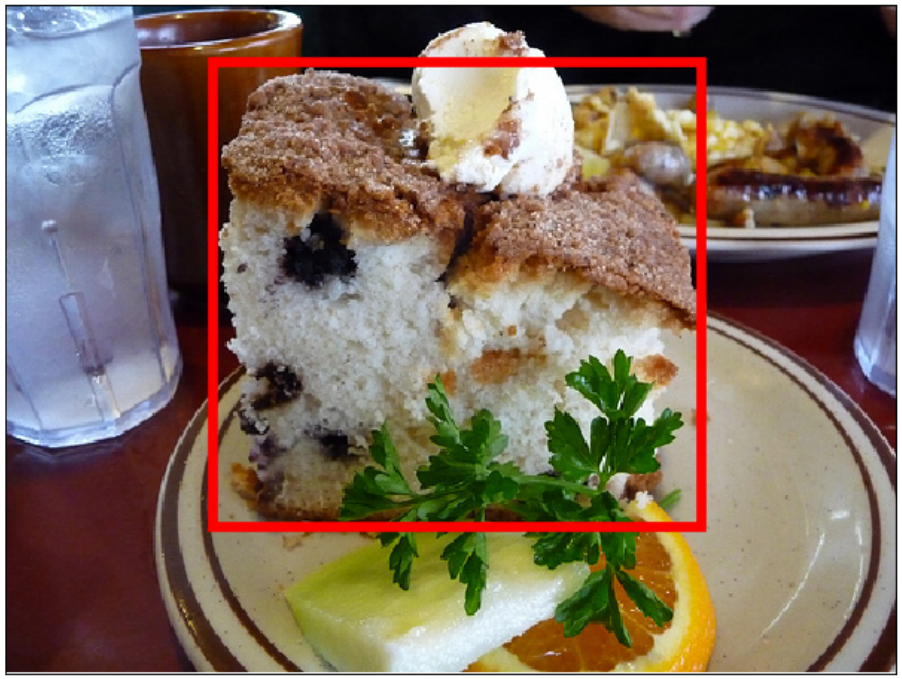
cake, food, bread, dessert, snacks, muffin, pastry

Object Naming in Real-World Images

[Silberer et al., 2020a, Silberer et al., 2020b]

Phenomenon of Object Naming: Existence of Preferred Name

VisualGenome+ManyNames

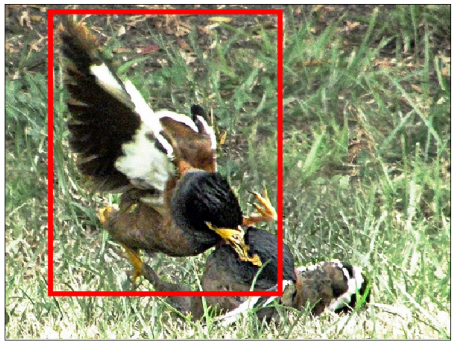


cake (53), food (19), bread (8), burger (6),
dessert (6), snacks (3), muffin (3), pastry (3)

Object Naming in Real-World Images

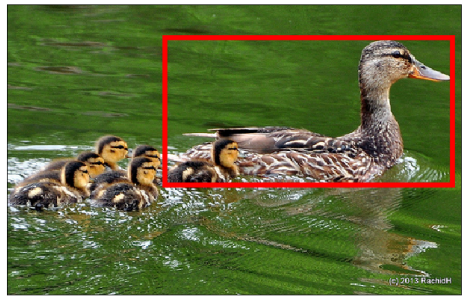
[Silberer et al., 2020a, Silberer et al., 2020b]

ManyNames



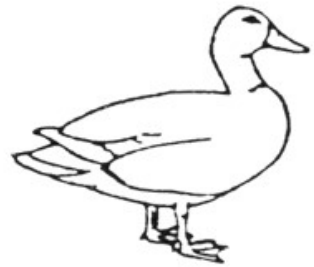
bird 27/duck 8

ManyNames



duck 33/bird 3

Snodgrass & Vanderwaart



duck (95%)

- Naming variation underlies certain factors
- Why analyse these factors?
  $\rightarrow$ linguistic analysis of human naming behaviour
  $\rightarrow$ model analysis & development
  $\Rightarrow$ Can models account for *human* object naming?

Object Naming in Real-World Images

[Silberer et al., 2020a, Silberer et al., 2020b]

- Do real-world objects have a canonical name?
- Are naming variants hierarchically related?

$\Leftarrow$ Elicit names from different speakers for the same instance

Object Naming in Real-World Images

[Silberer et al., 2020a, Silberer et al., 2020b]

ManyNames v1: Annotation Procedure

- 25,596 Images from VisualGenome (VG) [Krishna et al., 2017]
- Objects (images) from 7 domains, 52 WordNet synsets
- 36 names per object collected with Amazon Mechanical Turk

Name:

ManyNames

Domain	Top-1	Top-2
vehicles	train (954)	car (642)
food	pizza (518)	cake (261)
animals_plants	giraffe (915)	horse (822)
home	bed (888)	bench (714)
buildings	house (340)	bridge (274)
people	boy (853)	man (806)
clothing	shirt (904)	jacket (451)

Object Naming in Real-World Images

[Silberer et al., 2020a, Silberer et al., 2020b]

Do objects in real-world images have a canonical name?

VG: bridge

bridge (35)

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

Object Naming in Real-World Images

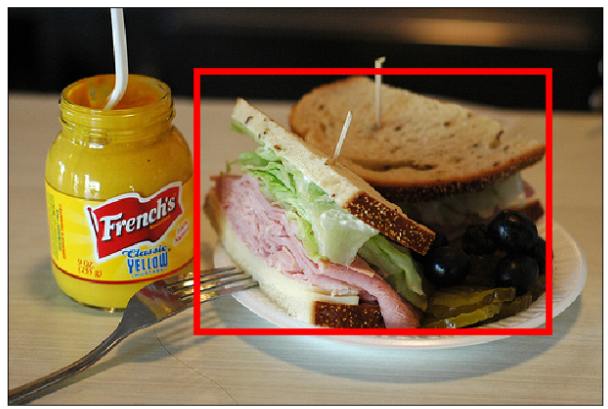
[Silberer et al., 2020a, Silberer et al., 2020b]

Do objects in real-world images have a canonical name?

Object Naming in Real-World Images

[Silberer et al., 2020a, Silberer et al., 2020b]

Do objects in real-world images have a canonical name?

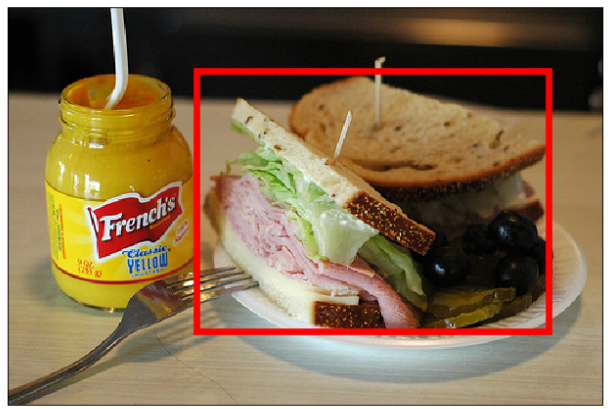


Object Naming in Real-World Images

[Silberer et al., 2020a, Silberer et al., 2020b]

Do objects in real-world images have a canonical name?

hotdog (14), food (7), bun (4), sandwich (3), bread (2) [banana (1)]



sandwich (34)

Object Naming in Real-World Images

[Silberer et al., 2020a, Silberer et al., 2020b]

Do objects in real-world images have a canonical name?

Are naming variations explainable by hierarchical relations?

VG: batter

boy (7), helmet (5), baseball player (4), player (4), man (3), child (3), batter (3), dress (2), kid (2)

pants (6), player (5), shoe (4), bat (4), person (4), legs (4), baseball player (3), hitter (2)

Object Naming in Real-World Images

[Silberer et al., 2020a, Silberer et al., 2020b]

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

- people, home, and buildings most susceptible to referential uncertainty
- vehicles and animals/plants trigger the least uncertainty, but most visual uncertainty, and linguistic errors

Object Naming in Real-World Images

[Silberer et al., 2020a, Silberer et al., 2020b]

Are naming variations explainable by hierarchical relations?

- 72.2% of naming variants are not due to different preferences in levels of specificity

VG: batter

boy (7), helmet (5), baseball player (4), player (4), man (3), child (3), batter (3), dress (2), kid (2)

pants (6), player (5), shoe (4), bat (4), person (4), legs (4), baseball player (3), hitter (2)

Object Naming in Real-World Images

[Silberer et al., 2020a, Silberer et al., 2020b]

Sources of Observed Naming Variation

Cross-classification:	conceptualization of alternative aspects of the same object (e.g. *boy-batter*, *dessert-toast*)
Metonymy:	logically related parts of an object stand in as its name (*sandwich-basket*, *bed-bed sheet*)
Conceptual disagreement	between speakers (*bed-bench*, *sandwich-hotdog*)

Object Naming in Real-World Images

[Silberer et al., 2020a, Silberer et al., 2020b]

Sources of Observed Naming Variation

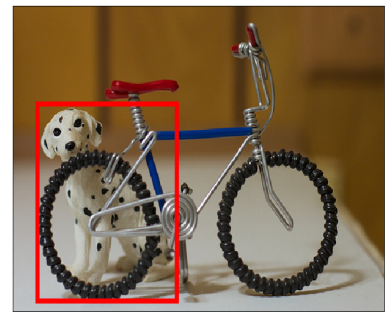
Cross-classification:	conceptualization of alternative aspects of the same object (e.g. *boy-batter*, *dessert-toast*)
Metonymy:	logically related parts of an object stand in as its name (*sandwich-basket*, *bed-bed sheet*)
Conceptual disagreement	between speakers (*bed-bench*, *sandwich-hotdog*)
Referential uncertainty	vs. Metonymy different objects for same bbox (e.g. *book-bed*, *pants-player*)
Visual uncertainty	vs. Conceptual disagreement

Object Naming in Real-World Images

[Silberer et al., 2020a, Silberer et al., 2020b]

C bear 16/polar bear 6/animal 3
I ball 5 | dog 3
E other-obj | inadequate

table 23/counter 5
food 6 | tabletop;desk
other-obj | singletons



dog 24/dalmatian 2
wheel 4 | ornament;toy
other-obj | singletons

Analysing Models on Human-Like Object Naming

[Silberer et al., 2020a, Silberer et al., 2020b]

Do Models Account for Human Object Naming in Images?

1. Can an object detector account for human object naming in images?
2. Does the model exhibit sensitivity to the interaction between domains and the visual characteristics of individual target objects that is similar to that of humans?

$\Rightarrow$ Do models predict a name that was the most frequent human response (n_{top}), valid but less frequent, or incorrect?

Analysing Models on Human-Like Object Naming

[Silberer et al., 2020a, Silberer et al., 2020b]

- Model: Bottom-Up [Anderson et al., 2018]
  representative L&V object naming model
- Target vocabulary: $1,253$ names
- Test set: $1,145$ images (objects)
- Human upper bound: evaluate all name tokens of response sets in MN v1

Analysing Models on Human-Like Object Naming

[Silberer et al., 2020a, Silberer et al., 2020b]

Diagnostic Evaluation Definition

Given: target object o_i and its name response set R_i

Does the predicted name $\hat{n}$ match ... ?

hit	n_{top}, the MN top name
same-obj	one of the valid, alternative names
other-obj	the name of another object in the bounding box
singleton	a name given once
inadequate	an invalid name
$\notin$	none of the names, $\hat{n} \notin R_i$

Analysing Models on Human-Like Object Naming

[Silberer et al., 2020a, Silberer et al., 2020b]

Model	correct			incorrect			
	n_{top}	same-obj	$\sum$	other-obj	inadequ.	singleton	$\notin$
Human	75.9	15.2	**91.1**	1.8	3.1	3.9	–
Bottom-Up	73.4	14.5	87.9	2.5	1.5	1.0	7.1

hit n_{top}, the MN top name

same-obj one of the valid, alternative names

other-obj the name of another object in the bounding box

singleton a name given once

inadequate an invalid name

$\notin$ none of the names, $\hat{n} \notin R_i$

Analysing Models on Human-Like Object Naming

[Silberer et al., 2020a, Silberer et al., 2020b]

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

- Model behaviour differs from humans on people, clothing, food
- Because model exhibits less variation, learns bias towards competing domain, ...

Analysing Models on Human-Like Object Naming

[Silberer et al., 2020a, Silberer et al., 2020b]

Examples

(f)

C cake 13, pie 3, cheesecake 2
I berry;dessert
E singletons
i *board* (∉)

(g)

suit 13/jacket 12/tuxedo 4
car;coat;hand;phone;tails
singletons
man (∉)

(h)

cake 24/food 5/bread 2
plate 4 | dessert
other-obj | singleton
bread (same-obj)

(i)

motorcycle 12/scooter 8/
wheel 2 | trike
other-obj | singleton
scooter (same-obj)

References to further Works

- Toward Human-Like Object Naming in Artificial Neural Systems https://baicsworkshop.github.io/pdf/BAICS_41.pdf
- Generating a Novel Dataset of Multimodal Referring Expressions https://aclanthology.org/W19-0507.pdf
- Chinese Whispers: A Multimodal Dataset for Embodied Language Grounding https://www.researchgate.net/publication/341294259_Chinese_Whispers_A_Multimodal_Dataset_for_Embodied_Language_Grounding
- CAT ManyNames: a New Dataset for Object Naming in Catalan https://aclanthology.org/2022.cogalex-1.4.pdf

References I

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., and Zhang, L. (2018).

Bottom-up and top-down attention for image captioning and visual question answering.
In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., Bernstein, M. S., and Fei-Fei, L. (2017).

Visual genome: Connecting language and vision using crowdsourced dense image annotations.
Int. J. Comput. Vision, 123(1):32–73.

Silberer, C., Zarrieß, S., and Boleda, G. (2020a).

Object naming in language and vision: A survey and a new dataset.
In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5792–5801.

Silberer, C., Zarrieß, S., Westera, M., and Boleda, G. (2020b).

Humans meet models on object naming: A new dataset and analysis.
In Scott, D., Bel, N., and Zong, C., editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1893–1905, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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.