code and data!

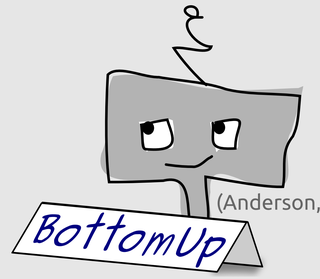

C&LING 2020 The 28th International Conference on Computational Linguistics

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

Carina Silberer, Sina Zarrieß, Matthijs Westera, Gemma Boleda

amore

upf. Universitat Pompeu Fabra Barcelona



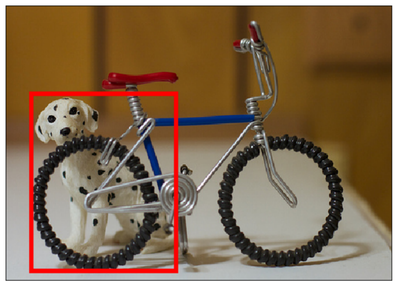

(Anderson, et al. 2018)

## Contributions:

1. **ManyNames v2**: A rich, reliable dataset of natural, human object naming behavior.
2. **Analysis** of representative object naming model, using ManyNames v2.

### Bear? Polar bear? Ball? Dog?

### Wheel? Dalmatian? Dog?



### Food? Table? Tabletop?

## The ManyNames dataset

**v1**
- naturalistic images (from VG).
- naming variation: 36 name tokens per object.
- 72K unique name-object pairs (for 25K objects).

Silberer et al. 2020 (LREC)

**New! v2**
- Filtered down to *consistent* name sets: only adequate names for the same object.
- 57K unique name-object pairs.
- Fine-grained inadequacy type annotations.

## How much *true* naming variation?

| domain | v1 names: | v1 top name: | v2 names: | v2 top name: |
|---|---|---|---|---|
| all | 2.9 | 75 % | 2.2 | 80 % |
| people | 4.3 | 59 % | 3.3 | 65 % |
| clothing | 3.2 | 70 % | 2.4 | 76 % |
| home | 3.1 | 72 % | 2.1 | 81 % |
| buildings | 3.0 | 74 % | 2.1 | 82 % |
| food | 2.9 | 76 % | 2.4 | 79 % |
| vehicles | 2.4 | 76 % | 2.1 | 78 % |
| animals/plants | 1.5 | 94 % | 1.3 | 95 % |

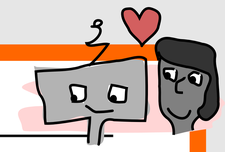
greatest variation reduction (mostly referential errors)

least variation reduction

the most 'tricky' cases

most visual errors here. Strong preference for 'basic level', even when uncertain!

## Model seems human-like...



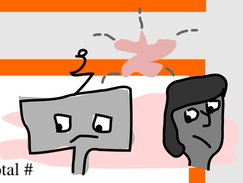
| Model | correct $n_{top}$ | correct same-object | correct all | incorrect other-object | incorrect inadequate | incorrect singleton | incorrect unobserved |
|---|---|---|---|---|---|---|---|
| Human | 75.9 | 15.2 | **91.1** | 1.8 | 3.1 | 3.9 | – |
| Bottom-Up | 73.4 | 14.5 | 87.9 | 2.5 | 1.5 | 1.0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

vs. 61% on original VG!

vs. =importance of taking naming variation into account!!!

similar error distribution

## ... but not in every domain



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

model is biased towards naming people

model worse; mostly referential errors

model exhibits less variation than humans for 'people'; more for 'vehicles'.

## ~~Take-home~~ Stay-home messages:

- Use *multi*-label evaluation on *naturalistic* images.
- Conduct fine-grained error analysis (types of error, & per domain).
- Rely on many annotators + simple verification step (*"same object? y/n"*)

## Acknowledgements

We thank the anonymous reviewers for their comments, and the AMT workers who participated in our crowdsourcing task. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 715154) and from the Catalan government (SGR 2017 1575). This paper reflects the authors' view only, and the EU is not responsible for any use that may be made of the information it contains.

erc European Research Council