Describe me an Auklet: Generating Grounded Perceptual Category Descriptions

Bill Noble* Nikolai Ilinykh*
Department of Philosophy, Linguistics and Theory of Science
Centre for Linguistic Theory and Studies in Probability (CLASP)
University of Gothenburg, Sweden
{name.surname}@gu.se

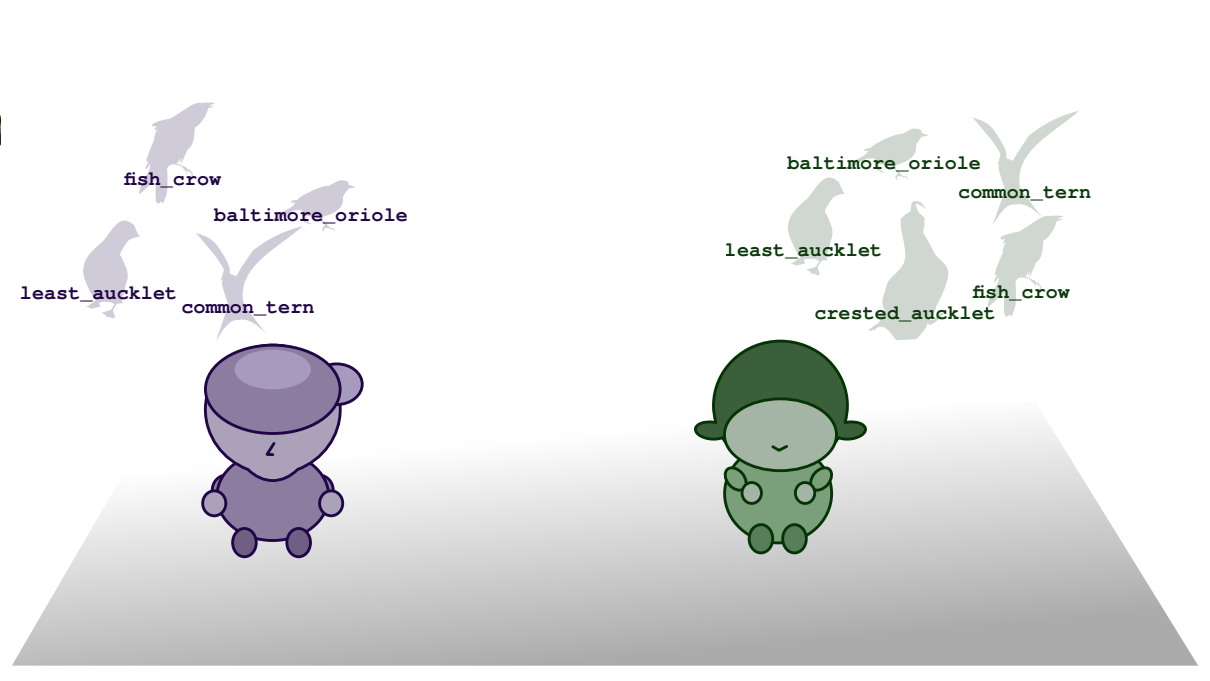
EMNLP – December, 2023

CLASP centre for linguistic theory and studies in probability

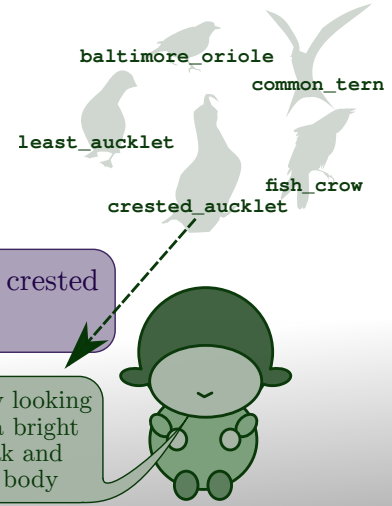
GÖTEBORGS UNIVERSITET

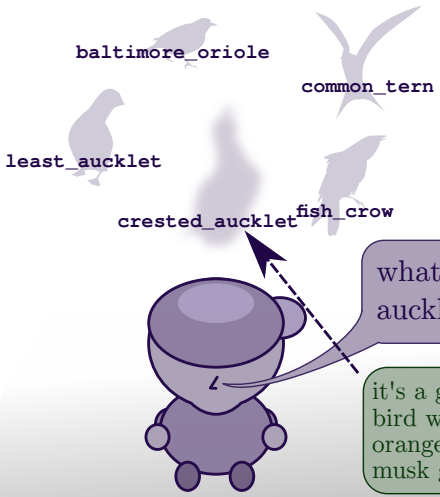
Grounding in context

- In NLP, grounding is often construed as a mapping between language and things in the world (Bernardi et al., 2016). E.g.,
  - image captioning
  - image retrieval
  - referring expression generation
- For humans, grounding is what makes language useful in for achieving our goals in a particular interactive context (Clark and Wilkes-Gibbs, 1986; Chandu et al., 2021; Larsson, 2018; Giulianelli, 2022).
- With *perceptual category description* we:
  1. make the interactive context explicit
  2. require grounding at an abstract level


fish_crow
baltimore_oriole
least_aucklet
common_tern
baltimore_oriole
common_tern
least_aucklet
fish_crow
crested_aucklet

Perceptual category description





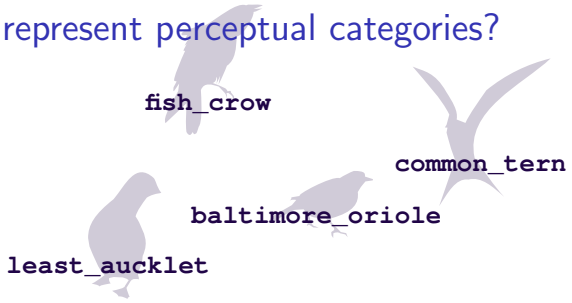
Perceptual Category Description: This paper

	Our setup
Domain	Bird species
Number of categories	200 (180 known / 20 unknown)
Perceptual modality	Photographs of individual birds[1] (50 per species) Collected from Flickr
Descriptions	English text descriptions[2] (500 per species) Annotators were given a diagram of bird bodyparts Instructed not to mention background or activity

Experimental focus: What *category-level representation* best support description generation in a model with a standard transformer decoder architecture?

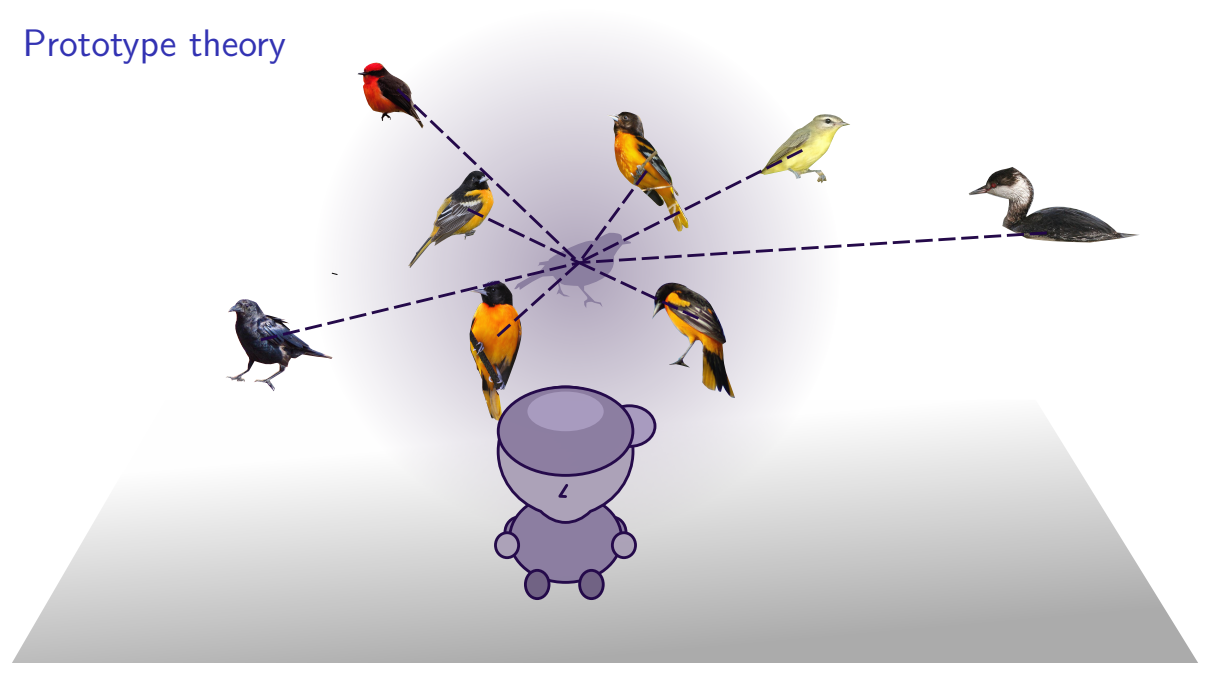
[1] Caltech-UCSD Birds-200-2011 (Wah et al., 2011)
[2] Reed et al. (2016)

How do humans represent perceptual categories?

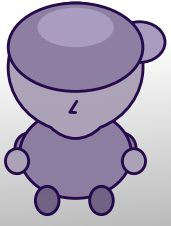


Prototype theory

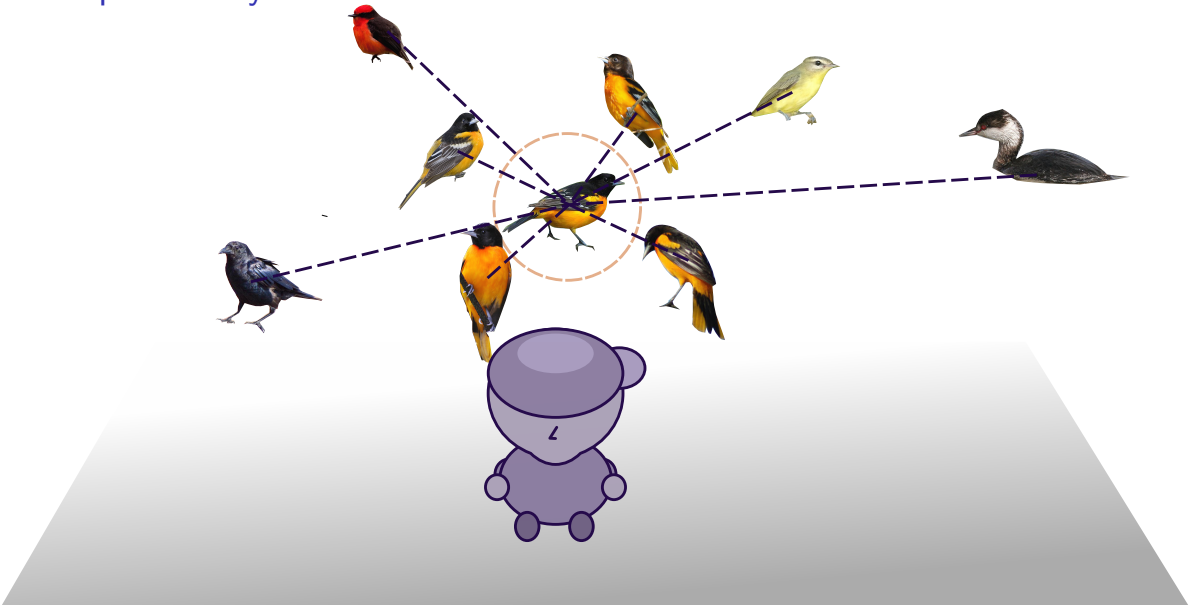
Prototype theory



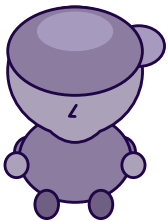
Exemplar theory



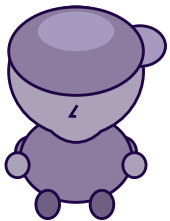
Exemplar theory

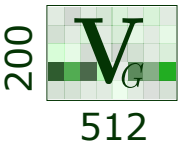


Models

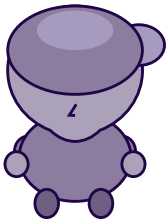


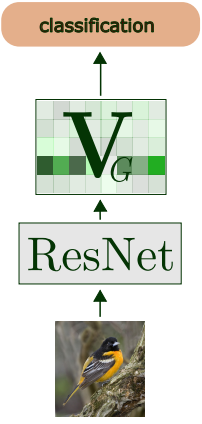
Generator model



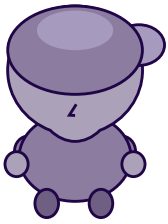


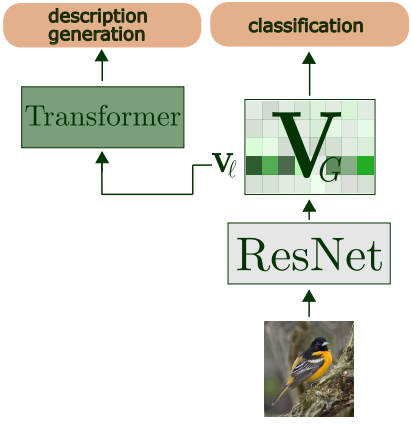
Generator model (training)



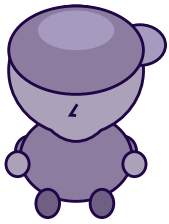


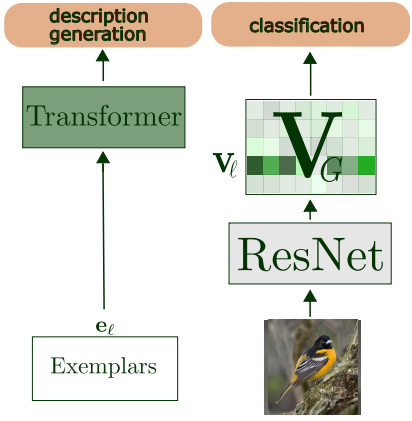
Prototype-based Generator model (training)



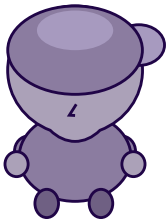


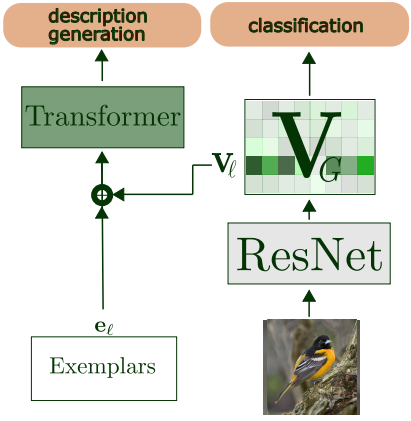
Exemplar-based Generator model (training)



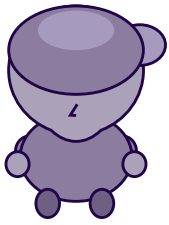


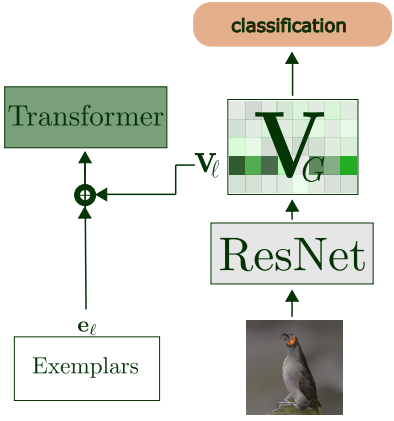
Hybrid Generator model (training)





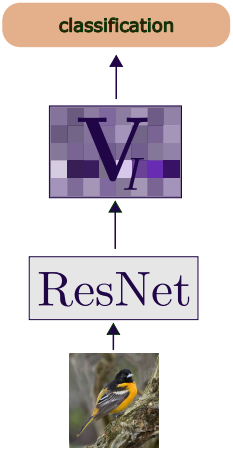
Generator model (training)



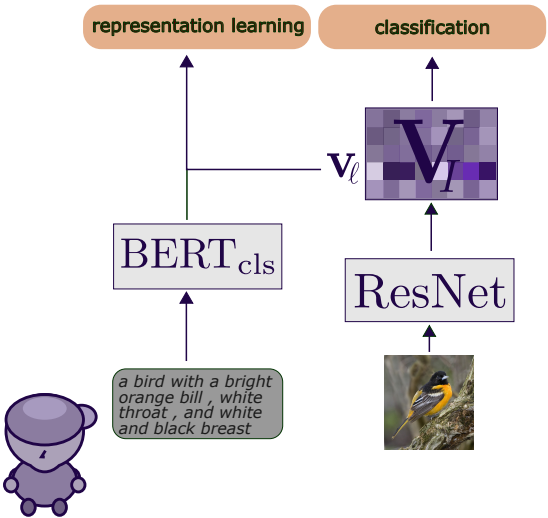


Interpreter model

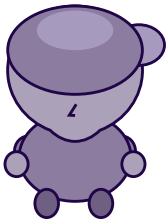
Interpreter model (training)

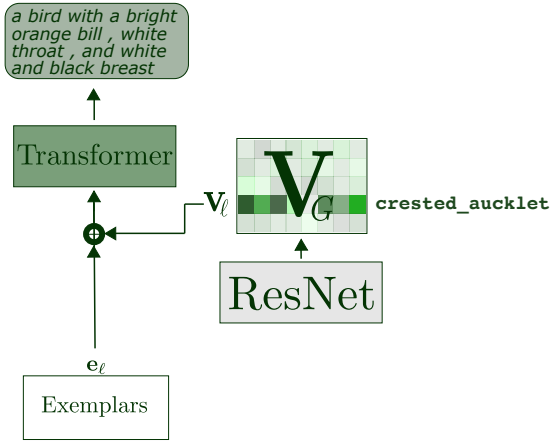


Interpreter model (training)

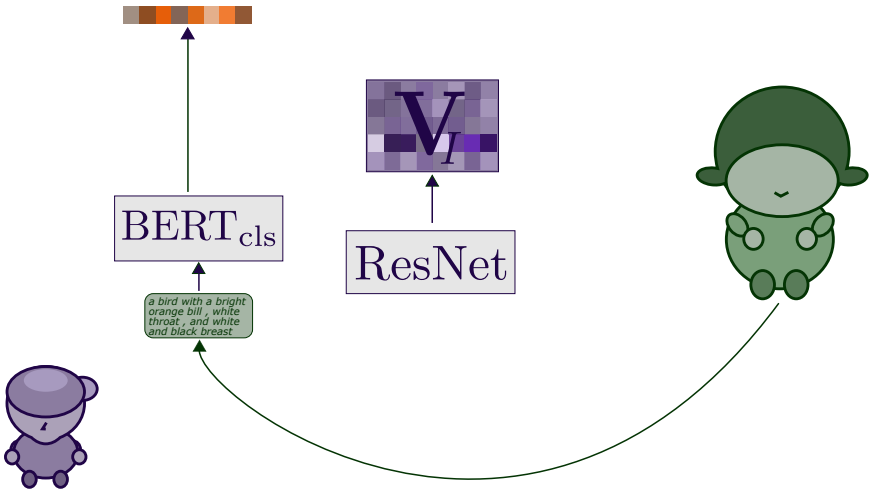


Generator model (inference)

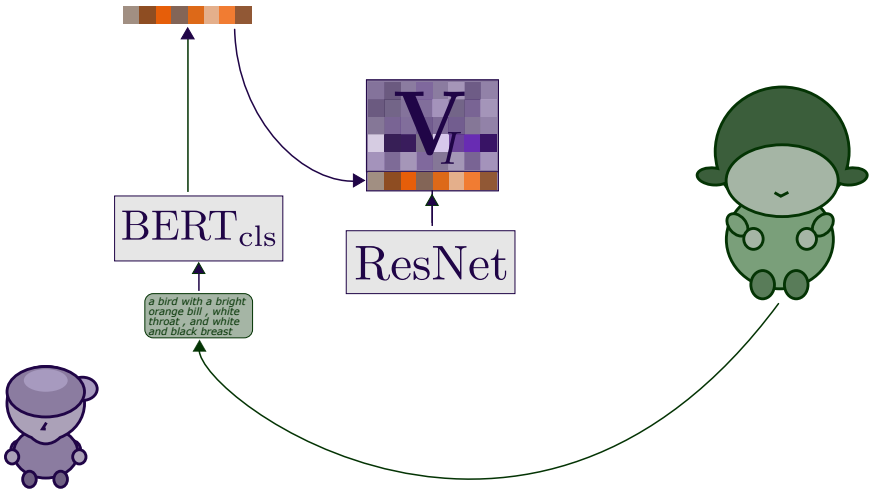




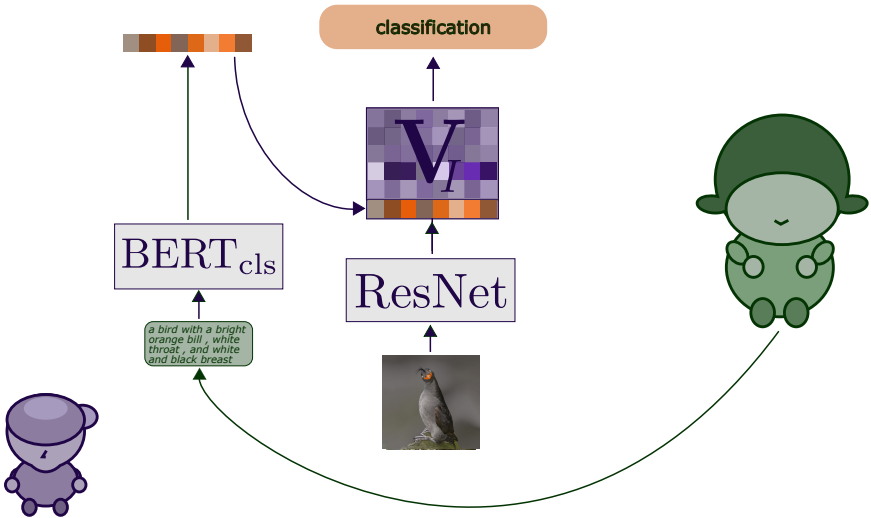
Interpreter model (zero-shot learning)



Interpreter model (zero-shot learning)



Interpreter model (zero-shot inference)



Interpreter results

teacher		CE loss	mean rank	acc@1	acc@5
ground truth	`seen`	2.50	5	36.4	75.1
	`unseen`	4.17	30	19.1	44.1
random baseline		5.30	100.5	0.5	2.5

Discriminativity

- The generator is trained to minimize cross-entropy vs. ground-truth descriptions, which may result in "safe" text
- We extract features, (noun phrases like "red wings") from the generated descriptions
- Based on the training set, we compute the *informativeness* of a feature x with respect to the class y:

$$\mathrm{disc}_Y(x) = \exp\big(H(Y = y) - H(Y = y|x)\big),$$

Generator results

class repr.	decoding	BLEU-1	BLEU-4	CIDEr	discriminativity mean	discriminativity max	accuracy @1	accuracy @5
both	beam	**.68**	**.55**	**1.83**	1.58	2.32	0.0	2.1
	nucleus	**.69**	.32	1.40	**5.22**	**12.48**	0.8	4.1
exem	beam	.64	**.58**	**1.92**	1.95	3.29	**8.6**	**25.1**
	nucleus	.65	.36	1.42	**5.10**	**12.07**	6.8	18.3
prot	beam	.61	**.55**	**1.80**	1.65	2.46	2.7	13.6
	nucleus	**.70**	.38	1.48	**5.51**	**13.14**	4.1	15.1

Conclusions

- Generating descriptions grounded in abstract perceptual concepts (as opposed to particular images) is difficult for standard neural architectures.
- Considering both intrinsic evaluation metrics (BLEU, etc.) and task-based metrics (communicative success) can give a more nuanced picture of grounding.
- Decoding strategy can have an impact on how grounded representations are expressed.

Takeaway messages:

- Groundedness is relative to communicative context.
- The task of perceptual category description is a method of investigating grounding at the level of abstract perceptual categories.

Future work:

- Diversify task parameters: *classification domain, modality*
- Experiment with larger pre-trained models. *How to deal with unseen class leakage?*
- A more interactive context: *Dialogue with feedback; reinforcement learning*

Thank you.

References I

Raffaella Bernardi, Ruket Çakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. 2016. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *J. Artif. Intell. Res.*, 55:409–442.

Khyathi Raghavi Chandu, Yonatan Bisk, and Alan W Black. 2021. Grounding 'grounding' in NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4283–4305, Online. Association for Computational Linguistics.

Herbert H. Clark and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition*, 22:1–39.

Mario Giulianelli. 2022. Towards pragmatic production strategies for natural language generation tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7978–7984, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Staffan Larsson. 2018. Grounding as a Side-Effect of Grounding. *Topics in Cognitive Science*, 10(2):389–408.

References II

Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. 2016. Learning Deep Representations of Fine-Grained Visual Descriptions. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 49–58, Las Vegas, NV, USA. IEEE.

C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. 2011. Caltech-ucsd birds 200. Technical Report CNS-TR-2011-001, California Institute of Technology.