Knowledge Supports Visual Language Grounding: A Case Study on Colour Terms

Simeon Schüz
Friedrich Schiller University Jena
simeon.schuez@uni-jena.de

Sina Zarrieß
Friedrich Schiller University Jena
sina.zarriess@uni-jena.de

Abstract

In human cognition, world knowledge supports the perception of object colours: knowing that trees are typically green helps to perceive their colour in certain contexts. We go beyond previous studies on colour terms using isolated colour swatches and study visual grounding of colour terms in realistic objects. Our models integrate processing of visual information and object-specific knowledge via hard-coded (late) or learned (early) fusion. We find that both models consistently outperform a bottom-up baseline that predicts colour terms solely from visual inputs, but show interesting differences when predicting atypical colours of so-called colour diagnostic objects. Our models also achieve promising results when tested on new object categories not seen during training.

1 Introduction

Research on human perception has shown that world knowledge supports the processing of sensory information (Mitterer et al., 2009; Ishizu, 2013). For instance, humans have been found to use their knowledge about typical colours of an object when perceiving an instance of that object, in order to compensate for, e.g., perceptually challenging illumination conditions and achieve colour constancy (Mitterer and de Ruiter, 2008; Witzel and Gegenfurtner, 2018). Thus, the visual perception of object colours can be thought of as leveraging *top-down* knowledge for *bottom-up* processing of sensory input, in accordance with traditional approaches in psychology (e.g. Colman, 2009). The integration of visual information and world knowledge in perception, however, is far from obvious, with views ranging from processing through bidirectionally connected bottom-up and top-down components to the assumption that visual and conceptual representations themselves are inseparably intertwined (Kubat et al., 2009).

Figure 1: Example object from VisualGenome with annotated colour attribute. The tree is described as "green", despite of challenging illumination conditions.

A lot of recent work in Language & Vision (L&V) has looked at grounding language in realistic sensory information, e.g. images of complex, real-world scenes and objects (Bernardi et al., 2016; Kafle and Kanan, 2017). In L&V, however, the use of top-down knowledge has mostly been discussed in the context of zero-shot or few-shot learning scenarios where few or no visual instances of a particular object category are available (Frome et al., 2013; Xian et al., 2018). [1]

We present a simple experiment on language grounding that highlights the great potential of top-down processing even for very common words with a lot of visual instances: we learn to ground colour terms in visual representations of real-world objects and show that model predictions improve strongly when incorporating prior knowledge and assumptions about the object itself. We investigate visual grounding of colour terms by combining bottom-up and top-down modeling components based on early and late fusion strategies, reflecting different interpretations about the integration of visual and conceptual information in human perception. We find that these strategies lead to differ-

[1] Note that in L&V, the term "top-down" has recently been used in a different way in the context of attention models where it refers to systems that selectively attend to the output of a certain layer (Anderson et al., 2018).

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 6536–6542

ent predictions, especially for atypical colours of objects that do have a strong tendency towards a certain colour.[2]

2 Related Work

Even recent work on colour terms has mostly been using artificial datasets with descriptions of isolated colour swatches that show a single hue, primarily examining effects of context and conversational adequacy in colour naming (Baumgaertner et al., 2012; Meo et al., 2014; McMahan and Stone, 2015; Monroe et al., 2016, 2017; Winn and Muresan, 2018). However, object colours bear a range of additional challenges for perception and grounding: (i) chromatic variation due to lighting and shading (Witzel and Gegenfurtner, 2018), (ii) effects of conventionalization as in e.g. *red hair* (Gärdenfors, 2004) and (iii) the inherent complexity of real-world objects (Witzel and Gegenfurtner, 2018), e.g. a tree with green leaves and a brown trunk is typically called green (see figure 1). In human cognition, several recalibration strategies support the constant perception of object colours given these challenges. In addition to bottom-up driven strategies like the chromatic adaption to situational sources of light, this also includes mechanisms such as the *Memory Colour Effect*: The automatic perception of canonical colours that accompanies the recognition of objects with characteristic hues (Olkkonen et al., 2008). Our aim in this work is to transfer knowledge-based recalibration mechanisms to the automatic classification of object colours.

Mojsilovic (2005) and Van de Weijer et al. (2007) propose pixelwise approaches for modeling colour naming in natural images, accounting for factors such as illumination and non-uniform object colours. Van de Weijer et al. (2007) assign colour terms as labels to colour values of individual pixels and then average over these labels to obtain a colour term for an image region. We use their model as one of our baselines in Section 4. However, they do not take into account object-specific colour tendencies. Zarrieß and Schlangen (2016) classify colour histograms for objects in real-world images. They train object-specific classifiers that recalibrate a bottom-up classifier, but only obtain a small improvement from recalibration. We implement a general top-down component that can be integrated with bottom-up processing in different ways.

3 Models

We focus on the effect of knowledge in language grounding and adopt a slightly idealized setting for modeling: we assume that the object type is available during training and testing. Following e.g. Snoek et al. (2005); Gunes and Piccardi (2008); Baltrusaitis et al. (2019), we distinguish early and late fusion as a way of integrating modeling components with different sources of information. Figure 2 illustrates our models, which we describe below.

BOTTOM-UP This component relies solely on sensory input and is implemented as a feed-forward network trained to predict colour terms from 3-dimensional RGB histograms (representing the polychromatic distribution of colour values in complex objects). The output layer has a softmax over the 11 basic colour terms (Berlin and Kay, 1969). For comparability, we adopt the architecture in Zarrieß and Schlangen (2016) (Input Layer: 512 nodes, Hidden layers with 240 and 24 nodes and ReLU activation, output layer: 11 nodes, Drop-Out: 0.2). We did not obtain improvements when testing other colour spaces. We also tried visual features extracted with a neural object recognizer (Simonyan and Zisserman, 2014) which only give a small improvement over colour histograms. Thus, in Section 4, we report results only for RGB histograms, as they are more transparent as representations and do not include any conceptual information on objects.

TOP-DOWN This component relies only on conceptual information about the object which consists of assignments of objects to object types and colour distributions for object types reflected in the data. Thus, this classifier predicts colour terms given only the object type, which is supposed to mimic the memory colour effect discussed in Section 2. We use (pre-trained) word embeddings for object types that are not fine-tuned during training. Hence, TOP-DOWN and the combined models can be tested on unseen object types. We use 100-dimensional pre-trained GloVe embeddings (Pennington et al., 2014). The embedding layer is followed by a hidden Layer (24 nodes, ReLU activation, drop-out set to 0.2).

LATE-FUSION In this approach, BOTTOM-UP and TOP-DOWN compute their classification de-

[2]Code and data for this project are available at: https://github.com/clause-jena/colour-term-grounding

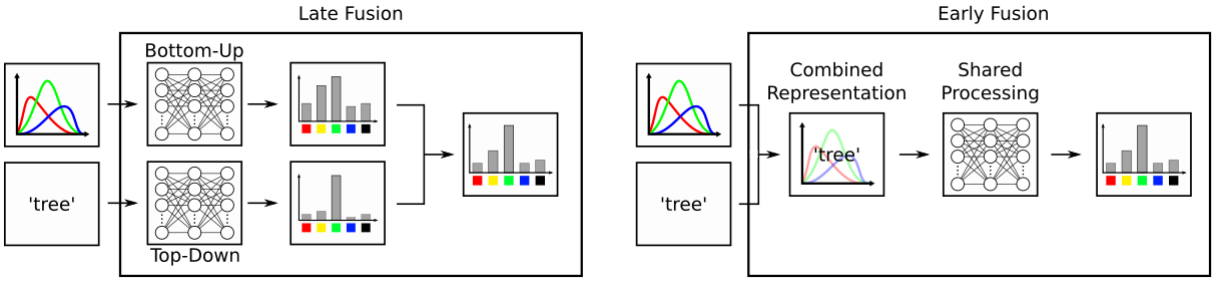

Figure 2: Late and Early Fusion

cisions independently. The output probability distributions are interpolated using a constant factor (which is set to 1 in our case), i.e. we simply calculate their arithmetic mean. Hence, the integration of visual and conceptual information is hard-coded.

EARLY-FUSION Object type embeddings are processed by a single Hidden Layer (24 nodes, ReLU activation, 0.2 drop out), concatenated with the visual input and then further processed by the network (2 Hidden Layers with 240 and 24 nodes, ReLU activation, 0.2 drop out). The classification decision is computed after this shared processing. The integration of both sources of information is therefore learned by the model.

4 Experiments

4.1 Set-up

Data We use VisualGenome (Krishna et al., 2016), which contains annotations and bounding boxes for 3.8M objects in more than 100K images. Roughly 2.8M object attributes are annotated, the most frequent being colour descriptions. We extracted all objects with at least one attribute among the basic colour terms *black, blue, brown, green, grey, orange, pink, purple, red, white, yellow*. Objects with multiple names were split up into distinct entries, basic colour terms were removed from object names. To counter VisualGenome's bias towards images of people (Krishna et al., 2016), we exclude objects with names that are hyponyms of *person*.[3] We compile our train and test data so that colours are evenly distributed, as we do not want the model to rely on biases in colour frequency.[4] For the development and evaluation sets, we use random under-sampling. To ensure training examples for less frequent object categories, 10K instances for each colour category are randomly picked from the original train set, with the possibility of objects being picked multiple times. In summary, 110k objects are used for training, 17523 for development and 9328 for evaluation. In Section 4.2, we report results for objects that occur at least 100 times in the data. For testing on unseen objects in Section 4.3, we use object types that occur at least 50 but less than 100 times with a colour attribute in VisualGenome (these are excluded from training).

Training We train for 25 epochs using RMSprop as optimizer and a learning rate of 0.001.

Evaluation We evaluate our models by measuring their accuracy both for the entire test set and for separate subsets of objects. In line with previous research in perceptual psychology (cf. Section 2), we distinguish **Colour Diagnostic Objects (CDOs)**, that are strongly associated with a specific *Memory Colour*, and **Colour Neutral Objects (CNOs)**, objects without a typical colour appearance. We expect the distinction between CDOs and CNOs to be reflected primarily in model predictions that involve the processing of conceptual object information. For CDOs, determining the respective Memory Colour could result in improved classification results, whereas this strategy is less promising for CNOs.

Manually identifying objects as CDOs or CNOs is hardly feasible when using large-scale data sets such as VisualGenome. We therefore decide on a quantitative basis whether object types exhibit characteristic colours, namely by means of the entropy of the colour term distribution of an object type. For each object type o, we determine p_c as the relative frequency of a colour c for all instances of the object. The entropy of an object's colour distribution is then calculated as

$$E_o = -\sum_{c \in C} p_c \, log_2 \, p_c$$

[3] Excluding e.g. "white person" as a case of a highly conventionalized colour.

[4] *white* has 290K, *purple* 10K instances in the data.

	All	CDO	CBO	CNO	CDO typ.	CDO atyp.
Pixelwise	38.5	50.4	32.5	41.0	58.6	26.6
BOTTOM-UP	45.0	54.0	36.5	50.4	62.7	28.9
TOP-DOWN	33.7	72.6	26.6	19.7	96.6	2.6
LATE-FUSION	52.1	71.7	43.4	51.1	94.0	6.9
EARLY-FUSION	51.4	74.0	43.7	48.5	94.0	15.7

Table 1: Accuracy in colour prediction for all seen object types (left); broken down for CDOs, CNOs, CBOs (middle), and for typical and atypical colours of CDOs (right)

		LATE-FUSION atypical	LATE-FUSION typical
Gold	atypical	33	272
	typical	53	834
		EARLY-FUSION atypical	EARLY-FUSION typical
Gold	atypical	69	236
	typical	53	834

Table 2: LATE-FUSION and EARLY-FUSION predictions for typical and atypical CDOs

where C is the set of basic colour terms. We use the 100 objects with the lowest entropy as CDOs, the 100 objects with the highest entropy as CNOs. In our data, objects such as *tree*, *carrot*, *jeans* and *refrigerator* are classified as CDOs. CNO examples include *balloon*, *umbrella*, *fish* and *butterfly*. We consider CDO instances whose colouring corresponds to their associated colour to be typical (e.g. objects annotated as *green tree*). Accordingly, CDOs that differ from their associated colour are considered atypical (e.g. *red tree*).

Some objects can neither be clearly identified as CDOs nor as CNOs. These include objects such as *stone* that often occur with specific colours (e.g. *grey*) but also with other colours (e.g. *brown, green*). To cover such cases we include **Colour Biased Objects (CBOs)** as a third group, determined as the 100 object types whose entropy is closest to the median of the data set.

Our test set contains a total of 1192 object instances categorized as CDOs (887 typical and 305 atypical) as well as 933 CBO and 1755 CNO instances.

Pixelwise Baseline For comparison, we report results of Van de Weijer et al. (2007)'s model on our data, that computes colour words for objects by classifying the individual pixels in the respective bounding box.

4.2 Results

Table 1 shows results for the separate model components and the fusion strategies. We note that BOTTOM-UP largely outperforms the pixelwise baseline. As expected, TOP-DOWN performs much worse than BOTTOM-UP on average, but achieves high accuracy on CDOs. We observe interesting differences between the fusion strategies:

LATE-FUSION This model generally performs better than BOTTOM-UP and TOP-DOWN separately. Moreover, the impact of the respective component on the combined result depends on the type of object: For CDOs, the model seems to generally predict the memory colour for these diagnostic objects computed by TOP-DOWN. For CBOs, there is still a clear improvement over BOTTOM-UP, whereas for CNOs the model mostly relies on BOTTOM-UP. Thus, even though the fusion is hard-coded, it achieves the desired flexible pattern for combining the components. However, LATE-FUSION does not perform well at predicting atypical colours of CDOs, see the right columns of Table 1. This suggests that, here, the prediction of object colours is only based on knowledge and, essentially, not visually grounded. This is unsatisfactory as these atypical colours for CDOs could be particularly salient in conversation (Tarenskeen et al., 2015).

EARLY-FUSION This fusion strategy generally improves the accuracy of BOTTOM-UP and TOP-DOWN in isolation. On average, it slightly underperforms LATE-FUSION, but obtains slightly better accuracy values for CDOs and CBOs than LATE-FUSION (Table 1). Table 2 illustrates that EARLY-FUSION recognizes atypical object colours slightly more often than LATE-FUSION. At the same time, the model achieves higher accuracy for atypical CDOs, indicating that it often predicts the correct object colour in these cases. For typical CDO colours, LATE-FUSION and EARLY-FUSION achieve the same accuracy.

Thus, EARLY-FUSION improves the prediction of atypical colours for CDOs as compared LATE-FUSION (exemplified in figure 3). But it still predicts canonical object colours too often and achieves a lower accuracy on atypical CDO colours than BOTTOM-UP. This indicates that early link-

object	% top colour	BOTTOM-UP Acc.	BOTTOM-UP % top prediction	EARLY-FUSION Acc.	EARLY-FUSION % top prediction
heater	94.12 (white)	0.0	35.29 (gray)	82.4	76.47 (white)
tablet	42.86 (black)	19.0	42.86 (blue)	61.9	57.14 (black)
wipers	94.12 (black)	35.3	29.41 (black)	70.6	76.47 (black)
room	54.55 (white)	18.2	31.82 (gray)	50.0	36.36 (white)
cherry	100.0 (red)	68.8	68.75 (red)	0.0	100.0 (green)
lime	100.0 (green)	68.4	68.42 (green)	5.3	94.74 (yellow)
dumpster	37.93 (green)	72.4	27.59 (blue)	10.3	75.86 (pink)
plank	57.14 (brown)	66.7	33.33 (brown)	4.8	90.48 (gray)

Table 3: Accuracy and top predicted colours for selected object types unseen during training. The top four objects obtain the highest improvements through early fusion, the bottom four objects decrease most with early fusion.

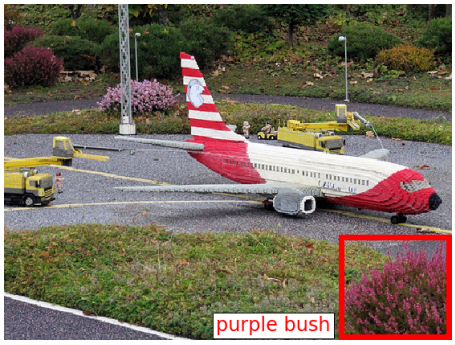

Figure 3: TOP-DOWN and LATE-FUSION predict the canonical colour for the depicted bush ("green"). BOTTOM-UP and EARLY-FUSION capture the annotated colour ("purple").

age and joint processing improves the integration of visual and conceptual information, at least for CDOs. It is, however, not a perfect solution to all problems identified: Even though the model learns to merge both sources of information, the bias for canonical colours is still too strong, and there remains a high dependence on non-sensory data.

4.3 Unseen Object Types

By using pre-trained embeddings, our models are able to handle object types that are unseen in the training set, via similarity to seen object types in the embedding space. For these objects, BOTTOM-UP and EARLY-FUSION achieve an overall accuracy of 37.8 and 31.9, respectively[5]. To provide more qualitative insights, Table 3 shows the top four and bottom four objects in terms of how much EARLY-FUSION improves over the BOTTOM-UP baseline. With *heater* and *wipers*, the top four objects include types with highly characteristic colours. EARLY-FUSION appears to correctly derive their object-specific colour tendencies from similarities to trained objects. In the lower four objects, all instances of *cherry* and *lime* share the same colour. Here, EARLY-FUSION also predominantly predicts a particular but incorrect colour, i.e. similarity in the off-the-shelf embedding space does not lead to good generalization for colour tendencies. This is particularly evident with *lime*: The prevailing prediction of *yellow* suggests that the (in this case, misleading) semantic similarity to the trained object type *lemon* is captured.

These findings support previous work on multimodal distributional semantics showing that off-the-shelf embeddings do not necessarily capture similarity with respect to visual attributes of objects (Silberer and Lapata, 2014).

5 Discussion and Conclusion

As in human perception, knowledge about typical object properties seems to be a valuable source of information for visual language grounding. Our fusion models clearly outperform a bottom-up baseline that relies solely on visual input. We also showed that the fusion architecture matters: the early integration of visual and conceptual information and their shared processing appears to be beneficial when colour diagnostic objects have atypical colours. However, even Early Fusion does not yet achieve a perfect balance between top-down and bottom-up processing. Future work should look at more complex fusion strategies, possibly coupled with bottom-up recalibration mechanisms (Zarrieß and Schlangen, 2016; Mojsilovic, 2005) to further enhance colour classification under difficult illumination conditions. Our experiment on objects unseen during training looks promising but can be extended towards a more general approach that interfaces colour prediction with object recognition.

[5]Note that these figures are not directly comparable with the results described above, since the instances for the individual colours are not evenly distributed in this set.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 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.

Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2019. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443.

Bert Baumgaertner, Raquel Fernandez, and Matthew Stone. 2012. Towards a flexible semantics: Colour terms in collaborative reference tasks. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics*, pages 80–84, Montréal, Canada. Association for Computational Linguistics.

Brent Berlin and Paul Kay. 1969. *Basic Color Terms: Their Universality and Evolution*. University of California Press, Berkeley and Los Angeles.

Raffaella Bernardi, Ruket Cakici, 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. *Journal of Artificial Intelligence Research*, 55:409–442.

Andrew M. Colman. 2009. *A dictionary of psychology*, 3 edition. Oxford paperback reference. Oxford University Press, Oxford.

Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc Aurelio Ranzato, and Tomas Mikolov. 2013. Devise: A deep visual-semantic embedding model. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2121–2129. Curran Associates, Inc.

Peter Gärdenfors. 2004. *Conceptual spaces: The Geometry of Thought*. A Bradford book. MIT Press, Cambridge, Mass. [u.a].

Hatice Gunes and Massimo Piccardi. 2008. Automatic temporal segment detection and affect recognition from face and body display. *IEEE transactions on systems, man, and cybernetics. Part B, Cybernetics : a publication of the IEEE Systems, Man, and Cybernetics Society*, 39:64–84.

Tomohiro Ishizu. 2013. Disambiguation of ambiguous figures in the brain. *Frontiers in Human Neuroscience*, 7:501.

Kushal Kafle and Christopher Kanan. 2017. Visual question answering: Datasets, algorithms, and future challenges. *Computer Vision and Image Understanding*, 163:3–20.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations.

Rony Kubat, Daniel Mirman, and Deb Roy. 2009. Semantic context effects on color categorization. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, pages 491–495. Cognitive Science Society.

Brian McMahan and Matthew Stone. 2015. A bayesian model of grounded color semantics. *Transactions of the Association for Computational Linguistics*, 3:103–115.

Timothy Meo, Brian McMahan, and Matthew Stone. 2014. Generating and resolving vague color references. In *Proceedings of the 18th Workshop Semantics and Pragmatics of Dialogue (SemDial)*, pages 107–115.

Holger Mitterer, Jörn M. Horschig, Jochen Müsseler, and Asifa Majid. 2009. The influence of memory on perception: it's not what things look like, it's what you call them. *Journal of experimental psychology. Learning, memory, and cognition*, 35 6:1557–62.

Holger Mitterer and Jan Peter de Ruiter. 2008. Recalibrating color categories using world knowledge. *Psychological Science*, 19(7):629–634.

Aleksandra Mojsilovic. 2005. A computational model for color naming and describing color composition of images. *IEEE Transactions on Image Processing*, 14(5):690–699.

Will Monroe, Noah D. Goodman, and Christopher Potts. 2016. Learning to generate compositional color descriptions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2243–2248, Austin, Texas. Association for Computational Linguistics.

Will Monroe, Robert X.D. Hawkins, Noah D. Goodman, and Christopher Potts. 2017. Colors in context: A pragmatic neural model for grounded language understanding. *Transactions of the Association for Computational Linguistics*, 5:325–338.

Maria Olkkonen, Thorsten Hansen, and Karl R. Gegenfurtner. 2008. Color appearance of familiar objects: Effects of object shape, texture, and illumination changes. *Journal of Vision*, 8(5):13–13.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Carina Silberer and Mirella Lapata. 2014. Learning grounded meaning representations with autoencoders. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 721–732.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Cees Snoek, Marcel Worring, and Arnold Smeulders. 2005. Early versus late fusion in semantic video analysis. pages 399–402.

S.L. Tarenskeen, M. Broersma, and B. Geurts. 2015. 'hand me the yellow stapler' or 'hand me the yellow dress': Colour overspecification depends on object category. In *Proceedings of the 19th Workshop on the Semantics and Pragmatics of Dialogue (SemDial)*, pages 140–148.

Joost Van de Weijer, Cordelia Schmid, and Jakob Verbeek. 2007. Learning color names from real-world images. In *CVPR 2007 - IEEE Conference on Computer Vision*, pages 1–8.

Olivia Winn and Smaranda Muresan. 2018. 'lighter' can still be dark: Modeling comparative color descriptions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 790–795, Melbourne, Australia. Association for Computational Linguistics.

Christoph Witzel and Karl R. Gegenfurtner. 2018. Color perception: Objects, constancy, and categories. *Annual Review of Vision Science*, 4(1):475–499.

Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. 2018. Zero-shot learning-a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*.

Sina Zarrieß and David Schlangen. 2016. Towards generating colour terms for referents in photographs: Prefer the expected or the unexpected? In *Proceedings of the 9th International Natural Language Generation conference*, pages 246–255, Edinburgh, UK. Association for Computational Linguistics.