Can You See the (Linguistic) Difference? Exploring Mass/Count Distinction in Vision

D. Addison Smith[1], Sandro Pezzelle[1],
Francesca Franzon[3], Chiara Zanini[4], Raffaella Bernardi[1,2]
[1]CIMeC, [2]DISI, University of Trento
[3]University of Padova
[4]University of Zürich
{first.last}@{unitn.it[1,2]|unipd.it[3]|uzh.ch[4]}

Abstract

This work explores the linguistic distinction between *count* and *mass* nouns in the visual modality. Since the former class typically refers to well-defined, countable objects, with the latter prototypically including less countable substances, we explore to which extent the linguistic distinction is grounded in the visual representations of the entities denoted by count/mass nouns. Using visual features extracted from a state-of-the-art Convolutional Neural Network (CNN), we show that the entities referred to as *mass* exhibit a lower variance both internally (i.e. intra-image) and externally (i.e. inter-image) compared to *count*. That is, various instances of substances are internally more homogeneous and externally more consistent to each other than are count. We compare variance across various CNN layers and show that it is indicative of the categorization when low-level features of the images are used, whereas any effect disappears when experimenting with higher-level, more abstract representations.

1 Introduction

The distinction between *mass* and *count* nouns is undoubtedly one of the most investigated topics in formal linguistics, at least since Cheng (1973) (see Fieder et al. (2014) for a brief review). At the simplest, descriptive level of analysis, mass nouns are usually paired with substances (e.g. water, flour, sand, etc.), cannot be inflected in plural form, and are preceded by indefinite determiners like 'some', 'much', 'a little', etc. In contrast, count nouns refer to isolable, well-defined objects (e.g. bicycle, house, tree, etc.), can take plural number, and are preceded by definite determiners like 'a/an', 'every', 'each', etc. Such a division is of course an oversimplification and leaves an uncertain zone in which semantic features do not map directly into morphosyntactic properties. In fact, nouns denoting the same referents may be used as mass in one language and as count in another (e.g. 'capelli' in Italian is countable, whereas 'hair' is mass). Moreover, nouns denoting aggregates like 'rice' or collections of semantically related objects such as 'furniture' or 'mail' may occur in mass contexts. While the status of these latter nouns is largely debated in literature (see Chierchia (1998); Doron and Müller (2010) for very opposite positions), substances in contrast are unanimously considered mass. Many theories have been proposed based either on the syntactic or the denotational aspects of the two groups (see among others Chomsky (1967); Allan (1980); Pelletier and Schubert (1989)). We do not enter into this debate and focus on a rather unexplored venue.

We investigate whether for the most prototypical cases, namely 'objects' for *count* and 'substances' for *mass*, the linguistic mass/count distinction is reflected in the perceptual properties of the referents. In support of a perceptual and conceptual, pre-linguistic difference between objects (usually denoted by count nouns in language) and substances (usually mass) are a number of studies reporting the ability of children to discriminate between them by relying solely on perceptual features of the entities, without using linguistic information (for a brief review see the introduction in Zanini et al. (2016)). To evaluate

our hypothesis, we employ a computational model trained to classify objects in images. We test whether *mass-substance* images are internally (i.e. among the various regions of the same image) more homogeneous, and externally (i.e. among the various instances of the same entity) more consistent compared to entities denoted by count nouns (see Figure 1). In other words, 'substances' should be distinguished from 'objects' by means of the lower *variance* of their visual features (somewhat similar to Kiela et al. (2015) in a lexical entailment detection task). Though similar with respect to shape, entities denoted by count nouns are likely to be very different with respect to many other low-level visual features (surface, texture, color, etc.). As a consequence, they would require higher-level operations to be recognized and classified as belonging to a particular entity class.

Figure 1: Left: images representing the count noun 'building'. Right: images representing the mass noun 'flour'. As can be noted, the former exhibits much more variability compared to the latter, both internally (i.e. among regions of the same image) and externally (i.e. among different images of the same entity).

Notable works have advanced the understanding of the perceptual cues linked to material recognition (Sharan (2009); Sharan et al. (2014)). To our knowledge, the present study is the first attempt to investigate the mass/count distinction in vision as linked to a linguistic perspective, namely the relationship between visual features and a grammaticalized opposition attested in language. The motivation is indeed different from the group of studies taking into account the distinction between *things* and *stuff* in the computer vision community (Tighe and Lazebnik (2010, 2013); Mottaghi et al. (2014); Caesar et al. (2016)). Caesar et al. (2016), for example, recently proposed an enriched version of the popular COCO dataset (Lin et al. (2014)) containing pixel-level annotation for *stuff* in addition to the source annotation for *things*. In this resource, however, the *stuff* class does not align with the *mass* category defined linguistically. To illustrate, it contains nouns like 'mirror', 'door', 'table', 'tree', 'mountain', and 'house', which are *count* nouns from a linguistic perspective. As a consequence, using existing resources is not feasible for our purposes. To test our hypothesis, we thus collect images representing objects and substances by relying on an existing lexical resource where countability annotation is available for English nouns. We then extract visual features from various layers of a pretrained state-of-the-art Convolutional Neural Network (CNN) and compute intra-image (i.e. between the various regions of an image) and inter-image (i.e. between different images depicting the same entity) variance at each layer. We show that visual features extracted from images depicting mass nouns exhibit a significantly lower intra-image variance compared to those representing count nouns, when computed at the early layers of the network (encoding low-level visual features). That is, mass nouns refer to simpler, more homogeneous entities compared to the more varied representations of count nouns. Moreover, at the early layers mass nouns are also more consistent between instances of the same object compared to count (i.e. lower inter-image variance). Consistent with our expectations, any effect disappears when experimenting with higher-level visual features extracted from the last layers of the network.

2 Dataset

To obtain mass/count categorization of nouns, and more specifically categorization of their respective senses, the Bochum English Countability Lexicon (BECL) (Kiss et al. (2016)) is used. This resource maps synsets within WordNet (Miller (1995)) to their respective countability classes, with noun senses annotated as either *mass*, *count*, *both*, or *neither* based on a series of syntactic patterns. The annotation occurs at the sense level, and a given noun can therefore have various senses belonging to distinct countability classes. Since our intention is to approach the matter from a vision perspective, we first check how many of the labeled synsets are available within ImageNet (Deng et al. (2009)), with an additional requirement that the images have available bounding box annotations. Among the available synsets are 36 *mass*, 58 *both*, and a remarkable 686 *count*.[1] This could well be a byproduct of the fact that count nouns are seemingly easier to annotate with bounding boxes given that they are discrete instances and are more often present in the foreground of an image. For this reason it could also be argued that more pictures are taken of count objects in general, which could explain the synset availability bias within ImageNet with regard to mass/count nouns.

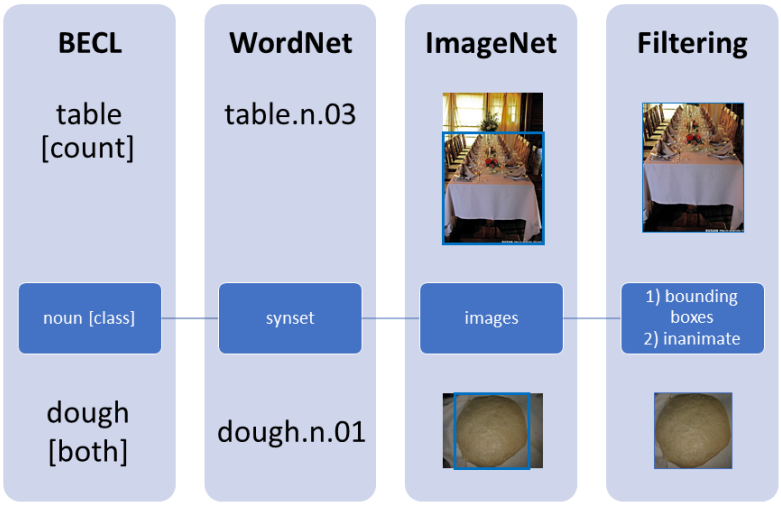

Figure 2: Various steps performed in building the dataset.

Investigation into the synsets annotated as *mass* reveals entities such as various sports ('soccer', 'basketball', etc.) whose corresponding images depict countable entities such as players or balls. We also encounter collective nouns such as 'equipment', 'furniture', 'luggage', 'housing', and 'artwork', which are merely collections of countable objects. The *both* category is therefore more suitable for our purposes given that the senses fit the prototypical idea of a mass noun as a substance. This *both* categorization in BECL is intuitive given that mass nouns can also be used in count contexts, i.e. 'two wines' which would refer either to two glasses (containers) of wine or perhaps to two different types of wine. In any case, nouns contained in this class ('flour', 'sugar', 'grain', etc.) are also viable from a vision perspective given their propensity for bounding box annotations. Since the *both* category captures mass-substance nouns, we henceforth refer to it simply as mass. Of these 58 mass noun senses none are animate, and so to avoid any possible confounding effects due to animacy we also constrain the countable objects to be inanimate, choosing the 58 most frequent where frequency is a BECL metric based on the Open American National Corpus (OANC). Images for the 58 + 58 synsets are downloaded and cropped according to bounding box annotations. Figure 2 illustrates the various steps followed to build the dataset, with descriptive statistics of the dataset reported in Table 1.

[1] We do not consider *neither* senses as they are very few and we do not find them useful for our purposes.

	#syns	#uniq_nouns	#imgs (avg)	#imgs (range)	OANC_freq (avg)	OANC_freq (range)
mass	58	56	214.66	64 - 705	112.6	10 - 447
count	58	53	303.93	60 - 1467	1435.16	33 - 4121

Table 1: Descriptive statistics of the dataset. From left to right, (1) number of synsets, (2) number of unique nouns among synsets, (3) average number of images per synset, (4) min, max number of images per synset, (5) average linguistic (OANC) frequency of the noun, (6) min, max frequency of the noun.

3 Experiments

To investigate the progression from the low-level, more concrete image features to the more abstract representations, we use a deep Convolutional Neural Network (CNN). This state-of-the-art CNN, namely the VGG-19 model (Simonyan and Zisserman (2014)), is pretrained on ImageNet ILSVRC data (Russakovsky et al. (2015)). VGG-19 consists of 5 blocks of convolutional layers (hence, *Conv*), each followed by a max pooling layer which extracts the most relevant features and hence reduces the dimensions of the feature vector. After the fifth convolutional block, 3 fully-connected layers (*fc*) are implemented. We evaluate 4 out of the 5 convolutional blocks (*Conv2*-*Conv5*) by extracting the outputs of the first and last layers for each block[2] and the output of the 3 fully-connected layers (*fc6*, *fc7*, and *fc8*). Convolutional layers are expected to capture low-level features (e.g. edges, texture, color, etc.) while the fully-connected layers compute abstract ones (see LeCun et al. (2015)). We check at which layer the *mass* and *count* synsets significantly differ with respect to their variance.

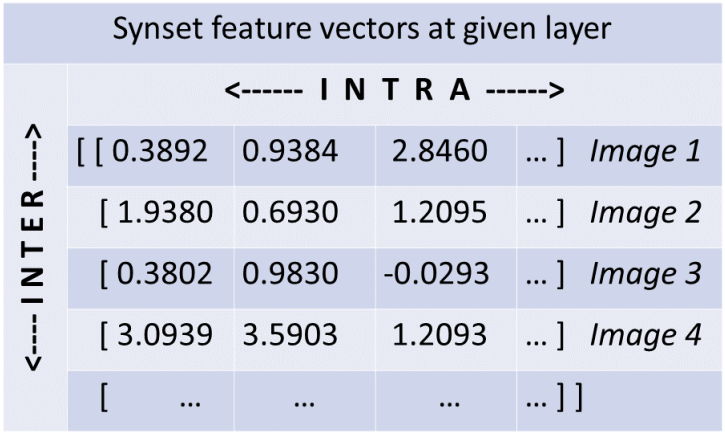

Figure 3: Toy representation of the two types of variance computed, i.e. *intra*- and *inter*-image.

Two types of variance are computed for all cropped images of a given synset: *intra*-vector (intra-image) variance and *inter*-vector (inter-image). See Figure 3 for a toy representation of both types of variance.

Intra-image After extracting and storing the feature vector for an image of a given synset at a given layer of the CNN, the variance of the feature vector is computed and subsequently averaged with the variances for all other images of the synset. This provides us with the mean *intra*-image variance, or the average variability within a single image of a given synset. This constitutes a measure of the relative homogeneity of the object, and picks up on the general complexity of the corresponding noun/sense.

Inter-image For the second type of variance, *inter*-vector variance, feature vectors for all images of a given synset are first extracted and stored from a given layer of the neural network. In this case, we calculate 'vertical' or column-wise variance among each individual dimension for all images of the synset, after which the dimension variances are averaged. This provides us with the *inter*-image variance, or the variability between distinct images of the same synset, which is a measure of the relative consistency

[2]Due to computational constraints, we do not consider the first *Conv1* block, which has approximately 3.2M dimensions.

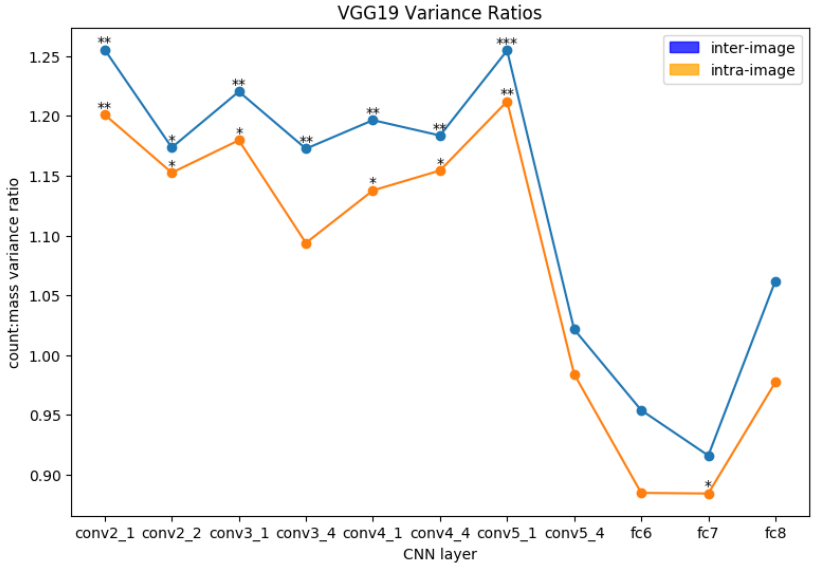

Figure 4: Difference between mass/count variance through the various layers of the network in both *intra-* (orange) and *inter-* (blue) settings. *** refers to a significant difference at p<.001, ** at p<.01, * at p<.05.

between instances of a given entity and its corresponding noun/sense.

Both types of variance are computed using the original, full-size vectors as extracted from the network.[3] That is, we do not employ any dimensionality reduction technique that could cause information loss affecting the variance values. To determine whether there is a significant difference between *mass* and *count* nouns, a two-tailed t-test is performed for each type of variance and for each layer of the CNN.

4 Results

We find both *intra*-image and *inter*-image variances to be significantly lower for *mass* nouns as compared to *count* nouns throughout all tested convolutional layers up until *Conv5_1*, with only one exception (*intra*-image variance in *Conv3_4*). From *Conv5_4*, in contrast, the difference becomes no longer significantly different, again with just one exception (*intra*-image variance in *fc7*).

Figure 4 shows this pattern of results obtained across the investigated layers. For visualization purposes, we plot the *ratio* between count and mass variance at each layer. As can be seen, this value is higher than 1 through the early layers, showing that count variance is higher than mass variance. Most importantly, within these layers (encoding low-level visual features) the difference in variance is overall very significant (as shown by the stars on the top of each 'node'). Throughout the convolutional blocks, the ratio indicating the difference between the two classes increases after the max pooling step is applied. This process ends at *Conv5_1*, when the more abstract visual features start to be computed by the network. Here, we observe quite a big drop in the count/mass ratio, showing that the two variances first become very similar and eventually 'change sign' (i.e. mass variance becomes higher than count). However, the difference in variance within these layers is generally not significant. Interestingly, at the last steps, especially at *fc8*, the ratio between the two variances stabilizes around 1, likely indicating that visual representations at this stage are abstract enough not to encode any information about the mass/count distinction. Zooming into the layers, *Conv5_1* turns out to be the layer where the difference in variance between mass/count synsets is highest for both settings (see Figure 5).

In Table 2 we report top-10 highest variance and bottom-10 lowest variance synsets obtained from

[3]Vector size ranges from 1.6M dimensions of *Conv2* to 1K dimensions of *fc8*.

Conv5_1 intra- variance		*Conv5_1 inter- variance*	
top-10	bottom-10	top-10	bottom-10
magazine_01 (c)	range_04 (c)	magazine_01 (c)	egg_yolk_01 (m)
salad_01 (m)	dough_01 (m)	shop_01 (c)	range_04 (c)
shop_01 (c)	mountain_01 (c)	salad_01 (m)	dough_01 (m)
church_02 (c)	mesa_01 (c)	machine_01 (c)	mountain_01 (c)
machine_01 (c)	flour_01 (m)	church_02 (c)	mesa_01 (c)
floor_02 (c)	milk_01 (m)	stage_03 (c)	milk_01 (m)
press_03 (c)	glacier_01 (m)	press_03 (c)	flour_01 (m)
stage_03 (c)	butter_01 (m)	floor_02 (c)	butter_01 (m)
pasta_01 (m)	egg_yolk_01 (m)	brunch_01 (m)	glacier_01 (m)
brunch_01 (m)	floor_04 (c)	building_01 (c)	sugar_01 (m)

Table 2: Synsets with highest (top-10) and lowest (bottom-10) variance in both *intra-* and *inter-* settings. Underlined synsets are those belonging to the lesser-represented class in a given column. The bottom-10 columns are presented starting from lowest variance and in ascending order.

this *Conv5_1* layer. As expected, most synsets in the top-10 columns belong to the count class (c), with synsets in the bottom-10 being mostly included in the mass class (m). Moreover, it can be noted that most of the synsets in the *intra-* setting also appear in the *inter-* setting, sometimes with an almost perfect alignment. Finally, by looking at the nouns that fall outside the expected pattern, we foresee some interesting cutting-edge cases (i.e. mass in top-10, count in bottom-10). 'Mountain' and 'range' (here with the sense of 'a series of hills or mountains'), for instance, are count nouns whose visual texture is intuitively homogeneous, as well as 'salad' and 'pasta' which are mass nouns referring to entities that consist *de facto* of many isolable parts, and thus vary more on average across instances.

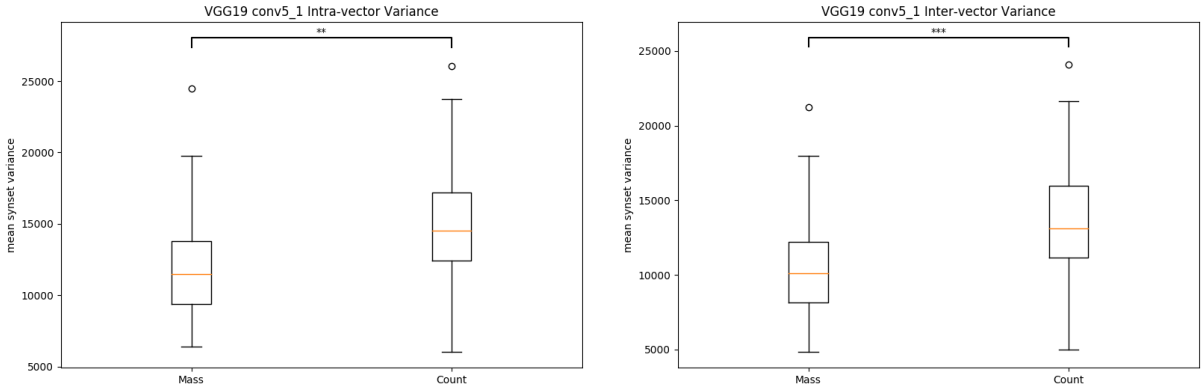

Figure 5: Boxplots reporting distribution of synset variance in both *intra-* (left) and *inter-* (right) setting for *Conv5_1* layer. Mass/count variance distribution is reliably different at $p<.01$ (left) and $p<.001$ (right).

5 Discussion

We show that mass-substance nouns have significantly lower intra- and inter-image variance than do count nouns, which is shown throughout the early convolutional layers of a state-of-the-art CNN. This could be useful in applications such as Visual Question Answering (VQA) or image caption generation, where a proper understanding of countability can lead to better responses and descriptions.

With that said, we can also see that there are cases which lie somewhere in the middle, exhibiting

visual properties belonging to the opposing mass/count class. Interestingly, count nouns which are labeled as 'stuff' in the *things* vs. *stuff* distinction, namely 'mountain', 'door', etc. (see Introduction), are found to behave more like mass in our experiments. More in general, there seems to be a visual *continuum* ranging from mass-substance all the way to count nouns. Similarly, recent studies in linguistics point to the fact that the distribution of nouns with respect to their syntactic contexts of occurrence is not consistent with a dichotomist division of the lexicon in two clear-cut classes of 'mass' and 'count' nouns (Zanini et al. (2016)). Also, metalinguistic judgments collected in various languages point to an interpretation of mass and count nouns as poles of a continuous distribution (Kulkarni et al. (2013)). Further investigations on the relation between the visual features of referents and cross-linguistic features are thus desirable.

Finally, the outcome showing lower variance for mass nouns in the inter-image setting might seem surprising, given that count nouns should overall refer to more well-defined objects, and thus more consistent in shape. However, this pattern of results is confirmed by literature dealing with object recognition in humans (Cichy et al. (2016)), where it is proposed that shape is a somewhat higher-level cognitive feature. In fact, when perceiving the real world, each visual experience of an entity is almost unique, due to changes in things such as orientation, lighting, and distance. This lack of invariance does not obstruct object recognition in the human observer, but its mechanisms are yet to be fully understood (DiCarlo et al. (2012)).

Acknowledgments

We kindly acknowledge Raquel Fernandez, Angeliki Lazaridou, Roberto Zamparelli, and Marco Marelli for their valuable insights and feedback. We are grateful to the Erasmus Mundus European Master in Language and Communication Technologies (EM LCT) for the scholarship provided to the first author. Moreover, we gratefully acknowledge the support of NVIDIA Corporation with the donation of the GPUs used in our research.

References

Allan, K. (1980). Nouns and countability. *Language*, 541–567.

Caesar, H., J. Uijlings, and V. Ferrari (2016). COCO-Stuff: Thing and Stuff Classes in Context. *arXiv preprint arXiv:1612.03716*.

Cheng, C. Y. (1973). Response to Moravcsik. In J. Hintikka, J. M. E. Moravcsik, and P. Suppes (Eds.), *Approaches to Natural Language*, pp. 286–288. Dordrecht: Reidel.

Chierchia, G. (1998). Reference to kinds across language. *Natural language semantics 6*(4), 339–405.

Chomsky, N. (Ed.) (1967). *Aspects of the Theory of Syntax*. Cambridge. Mass: MIT Press.

Cichy, R. M., A. Khosla, D. Pantazis, A. Torralba, and A. Oliva (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific reports 6*.

Deng, J., W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei (2009). Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 248–255. IEEE.

DiCarlo, J. J., D. Zoccolan, and N. C. Rust (2012). How does the brain solve visual object recognition? *Neuron 73*(3), 415–434.

Doron, E. and A. Müller (2010). The cognitive basis of the mass-count distinction: Evidence from bare nouns. *Abstract. The Hebrew University of Jerusalem and University of Sao Paolo*.

Fieder, N., L. Nickels, and B. Biedermann (2014). Representation and processing of mass and count nouns: a review. *Frontiers in psychology 5*.

Kiela, D., L. Rimell, I. Vulic, and S. Clark (2015). Exploiting image generality for lexical entailment detection. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL 2015)*, pp. 119–124. ACL.

Kiss, T., F. J. Pelletier, H. Husic, J. M. Poppek, and R. N. Simunic (2016). A Sense-Based Lexicon for Count and Mass Expressions: The Bochum English Countability Lexicon. In *Proceedings of LREC*.

Kulkarni, R., S. Rothstein, and A. Treves (2013). A statistical investigation into the cross-linguistic distribution of mass and count nouns: Morphosyntactic and semantic perspectives. *Biolinguistics 7*, 132–168.

LeCun, Y., Y. Bengio, and G. Hinton (2015). Deep learning. *Nature 521*(7553), 436–444.

Lin, T.-Y., M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick (2014). Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755. Springer.

Miller, G. A. (1995). Wordnet: a lexical database for English. *Communications of the ACM 38*(11), 39–41.

Mottaghi, R., X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille (2014). The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 891–898.

Pelletier, F. J. and L. K. Schubert (1989). Mass expressions. In *Handbook of philosophical logic*, pp. 327–407. Springer.

Russakovsky, O., J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision 115*(3), 211–252.

Sharan, L. (2009). *The perception of material qualities in real-world images*. Ph. D. thesis, Massachusetts Institute of Technology.

Sharan, L., R. Rosenholtz, and E. H. Adelson (2014). Accuracy and speed of material categorization in real-world images. *Journal of vision 14*(9), 12–12.

Simonyan, K. and A. Zisserman (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Tighe, J. and S. Lazebnik (2010). Superparsing: scalable nonparametric image parsing with superpixels. *Computer Vision–ECCV 2010*, 352–365.

Tighe, J. and S. Lazebnik (2013). Superparsing. *International Journal of Computer Vision 101*(2), 329–349.

Zanini, C., S. Benavides-Varela, R. Lorusso, and F. Franzon (2016). Mass is more: The conceiving of (un) countability and its encoding into language in 5-year-old-children. *Psychonomic Bulletin & Review* (24), 1330–1340.