An exploration of visual cues related to countability: two computer vision experiments.

Francesca Franzon[1,2], Chiara Zanini[2,3], D. Addison Smith[4],
Raffaella Bernardi[4], Sandro Pezzelle[4]

[1]SISSA, [2]University of Padova, [3]University of Zürich, [4]University of Trento
[1]ffranzon@sissa.it {first.last}@{[3]uzh.ch|[4]unitn.it}

1 Introduction

Some linguistic theories describe the mass-count distinction as the lexical encoding of a binary semantic opposition mapping some physical properties of the referential entities, namely the fact of denoting a substance or an object (Cheng, 1973; Jackendoff, 1991). The lexical feature related to countability would rule the possibility for nouns to occur in different syntactic contexts, so that nouns denoting objects should not occur in mass contexts (**some chair*) and nouns denoting substances should not occur in count contexts (**a butter*) unless taking into account semantic shift operations.

However, data on the distribution of nouns collected in different languages point to the fact that most nouns are accepted as grammatical and occur frequently in both syntactic contexts. Nouns do not gather in two separate groups and are spread along a continuum instead, with a slight bias in favour of count contexts (Katz & Zamparelli, 2012; Kulkarni et al., 2013; Schielen & Spranger, 2006; Zanini et al., 2017). These data are more consistent with theories that point to the role of the linguistic context of occurrence in denoting the countability of nominal expressions (Allen 1980; Pelletier 2012; Rothstein 2010). The syntactic context of occurrence is crucial to disambiguate differences in meaning even related to the same lexical noun, for example in oppositions like *a pizza* vs. *some pizza*, where the former phrase denotes the referential entity as bounded, and the second does not entail a reference to boundaries.

The grammaticalization of a reference to boundaries may reflect the importance that these have at the cognitive level, as it has been theorized for salient information concerning other extra-linguistic cognitive domains (Strickland, 2016; Franzon et al, forthcoming). Indeed, boundaries are crucial to recognize and conceive objects. It follows that encoding a reference to the boundaries of an object when denoting it in the language is often very pertinent. Yet, in our experience of the world, even substances are perceived within boundaries. Nevertheless, encoding a reference to boundaries when denoting substances may be not as pertinent as it is when denoting objects. Boundaries and shape are some of the features which play a pivotal role in the conceiving of entities as things or as stuff. In this regard, the arbitrariness of entities' structure, its repetition and its regularity are equally important cues (Prasada et al. 2002; Huntley-Fenner, Carey & Solimando 2002). In turn, the conceiving of these features relies to the processing and integration of lower level perceptual cues (Cichy et al., 2016).

In order to investigate the progression from the low-level, more concrete image features (e.g. edges, texture, color, etc.) to the more abstract representations, we designed two computer vision experiments exploiting a deep Convolutional Neural Network (CNN), the former evaluating the variance of the visual vectors, the latter aimed at classifying them with a Support Vector Machine (SVM) technique.

Previous works in computer vision targeting similar issues have focused on the distinction between things and stuff (Caesar et al., 2016; Tighe and Lazebnik, 2010, 2013; Mottaghi et al., 2014) and on materials vs object recognition (Fleming, 2013; Schwartz & Nishino, 2015; Sharan et al., 2009, 2014). To our knowledge no study has explored the relation between the visual properties of the images and the fact that their corresponding nouns are preferentially labeled as substances or as objects. In the present work, we focused on the role of visual cues of the entities of the referential world that may predict the pertinence for a noun to be used as mass or as count.

2 Experiments

2.1 Dataset

To obtain mass/count categorization of nouns, and more specifically categorization of their respective senses, we used the Bochum English Countability Lexicon (BECL) (Kiss et al., 2016). 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. A given noun can therefore have various senses belonging to the syntactic contexts in which it can occur. Since our intention is to approach the matter from a vision perspective, we first checked 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 there are 36 *mass*, 58 *both*, and, remarkably 686 *count*.[1]

This uneven distribution could be a byproduct of the fact that nouns denoting objects 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. Notably, this preeminence seems to reflect the bias for countability reported in the linguistic distribution observed in corpora.

Nouns denoting sports or activities that do not map to a concrete referent (such as *soccer*), as well as collective nouns as *luggage*, were not included although tagged as *mass*. For our experiments, nouns denoting substances were assingned to the *mass* class. Notably, in BECL nouns of substances are mainly categorized as *both*, as the annotation took into account also the sense in which they are used to define bounded entities (*a coffee*). Nouns contained in this class (*flour, sugar, grain*, etc.) are also viable from a vision perspective given their propensity for bounding box annotations.

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. The left panel on Figure 1 illustrates the various steps followed to build the dataset, with descriptive statistics of the dataset reported in Table 1.

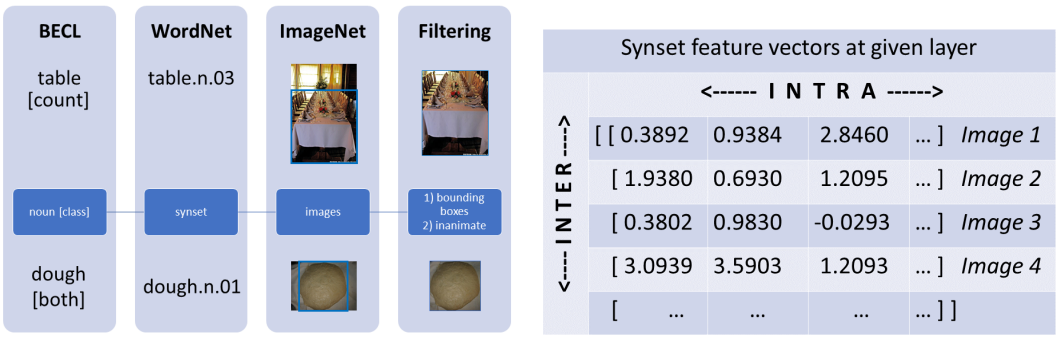

Figure 1: Left: The various steps performed in building the dataset. Right: Toy representation of the two types of *variance* computed, i.e. *intra*- and *inter*-image.

2.2 Visual variance of images

To investigate the progression from the low-level image features (e.g. edges, texture, color, etc.) to the more abstract representations, we used a deep Convolutional Neural Network (CNN), namely the state-of-the-art VGG-19 model (Simonyan & Zisserman, 2014), which 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

[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	1556.17	624 – 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.

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. 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 the right panel on Figure 1 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 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. 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 2 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

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

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

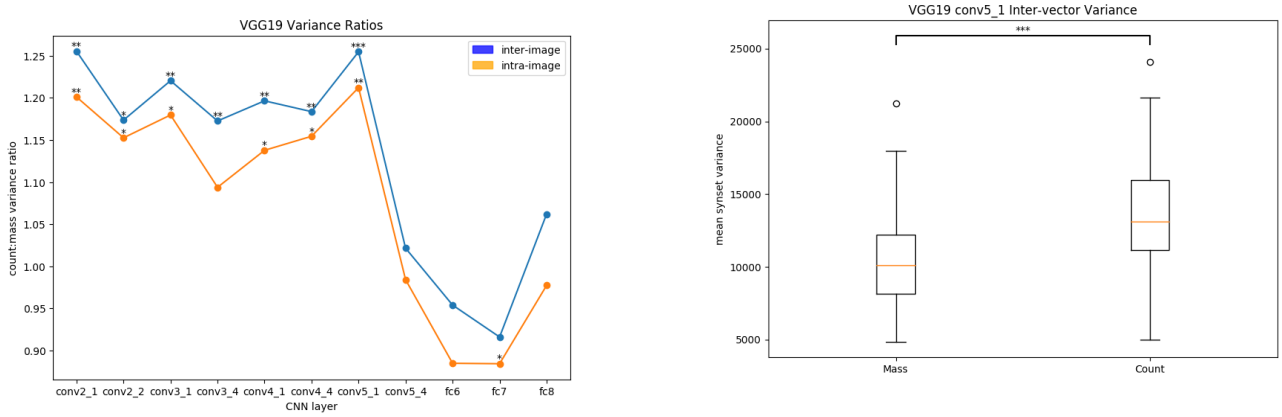

Figure 2: Left: Difference between mass/count variance through the various layers of the network in both *intra-* (orange) and *inter-* (blue) settings. Right: Boxplot reporting distribution of synsets *inter-* variance for *Conv5_1*. *** refers to a significant difference at p<.001, ** at p<.01, * at p<.05.

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. We calculated the top-10 highest variance and bottom-10 lowest variance synsets obtained from this *Conv5_1* layer.

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. The bottom-10 columns are presented starting from lowest variance and in ascending order.

As expected, most synsets in the top-10 columns belong to the count class (c) and synsets in the bottom-10 mostly belong to 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.

2.3 SVM classification

The first experiment showed us that mass and count nouns significantly differ with regard to visual variance both within and between images. Given these results, a plausible curiosity is whether or not it is possible to *classify* entities in images as either mass or count based on the very same feature vectors from which variance was computed. To this end, Support Vector Machine (SVM) classifiers are trained and tested [4] to evaluate the hypothesis that indeed the mass versus count distinction should

[4]Linear kernel employed via sklearn.svm.LinearSVC in Python 2.7.

be inherent in the visual vectors and yield overall high classification accuracies (while varying slightly depending on layer). SVMs manipulate features into *support vectors* which map datapoints into a high-dimensional space with a hyperplane separating the two classes to the greatest extent possible.

The previous 58 mass and 58 count synsets were randomly split 80%/20%, hence giving $47 + 47$ training synsets and $11 + 11$ testing synsets. This random split was consistently repeated 10 times for each normalization and layer combination, after which resulting accuracies were averaged for each. This provided an average training datapoint size of 24,600 and average testing datapoints of 5,682 (or roughly an 81%/19% split in the end). Layers selected for testing include each of the three fully-connected layers from VGG-19 (*fc6*, *fc7*, *fc8*), partly due to their reduced dimensionality,[5] as well as *Conv5_1* which showed the most significant difference in visual variance from the previous experiment.

To this end, we trained a Support Vector Machine (SVM) classifier to distinguish images within the two classes. We then evaluate the SVM on test images from novel synsets comparing various normalization techniques. The system achieves high classification accuracies overall, with a maximum accuracy of 87.9%. To note, the data was split among *synsets* in order to make the classification a bit more challenging, i.e. the split could have also been a simple 80%/20% split among *all* datapoints, ignoring their respective synset labels. However, this would mean the classifier would see examples of all synset classes in training, which could make the classification task in testing significantly easier. Instead, the decision was made to test the mass/count classification on *novel* synsets, or ones that have not been previously seen by the system.

CNN layer:	conv5_1	fc6	fc7	fc8
no_norm	**0.835**	0.817	0.821	0.828
norm_0	0.844	0.812	0.806	0.856
norm_1	**0.861**	**0.873**	**0.872**	**0.879**

Table 3: Overall SVM accuracies trained to distinguish mass versus count from visual vectors. Various normalization techniques are explored among various CNN layers, and each entry is an average of 10 random train/test data splits.

Two different normalization techniques[6] of the feature vectors were tested: *norm_0*, which normalizes by *dimension*, and the default *norm_1* that normalizes by *sample*. This is similar to the inter- and intra-vector variance, respectively, in that norm_0 normalizes a single *dimension* at a time (among all samples), whereas norm_1 normalizes a single *sample* at a time (independently of all other samples). This is compared against SVM classification without any prior normalization of feature vectors (*no_norm*).

It is clear from the data that the standard practice of normalizing vectors by sample (norm_1) provides superior results. The accuracies presented in Table 3 show the proportion of correctly-classified mass and count feature vectors. The best results are obtained from the final fully-connected layer, *fc8*, which makes sense given that this is the final layer before the CNN outputs the type of object in a given image. Since the visual information has been made abstract at this level, the network is able to generalize well about the countability of the respective objects as well. Also to note, *Conv5_1* has the highest classification accuracy for no_norm which is expected from the results of Experiment I: since the visual vectors vary the most between mass and count nouns at this layer, it follows that these discrepancies can be exploited for classification in this task. Furthermore, the normalization techniques can be said to confound the ability of the classifier to capitalize on variance in the visual features, which gives the upper hand to the fully-connected layers in the normalized settings.

3 Discussion and conclusions

Our findings point to the fact that pictures tagged with mass-substance nouns show significantly lower intra- and inter-image variance than do pictures tagged with count nouns. Such a result may

[5] 4096, 4096, and 1000 dimensions respectively.

[6] *l2* 'soft' normalization via sklearn.preprocessing.normalize utilizing default axis 1 and also axis 0.

seem counter-intuitive when approaching the issue following a classical linguistic perspective, since objects have often been described as having well-defined, non-arbitrary and consistent shapes. This consistency is though a result of a high level cognitive processing. In fact, literature on perception consistently agrees that an object is recognized although the lack of invariance of its low level visual features, as it is a higher level process. That explains the ability to recognize a specific object independently of its positional, scaling or illumination variability, but also the ability to recognize - and accordingly label - different entities as belonging to the same category, like *car, face, dog*(for a perspective on this, see Di Carlo et al., 2010). The complexity of the count class is consistent with this, whereas the relative homogeneity of the mass nouns is a result that may lead to new explorations in the relation between perception, cognition and language. Notably, we selected and analyzed only pictures representing almost pure mass and pure count nouns in order to magnify the difference. Yet, the obtained pattern is not so clear-cut. Interestingly enough, many pictures referring to objects, such as (*mountain*)or (*door*) behave like mass in our experiments. This seems consistent with the data reported in several rating and corpus linguistic studies in which mass and count nouns are found to be spread along a continuum rather than distributed in two poles (e.g. Kulkarni et al. 2013; see the Introduction). From the linguistic point of view, these data fit better linguistic theories claiming for a non lexical encoding of the countability features into language. From these starting observations, also further investigations tackling different languages or different corpora are desirable.

References

Allan, K. (1980). Nouns and countability. *Language* 56, 541-567.

Caesar, H., Uijlings, J., & Ferrari, V. (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*, 286–288. Dordrecht: Reidel.

Cichy, R.M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Deep Neural Networks explain spatio-temporal dynamics of visual scene and object processing. *Journal of Vision, 16(12)*, 371-371.

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

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

Fleming, R. W. (2014). Visual perception of materials and their properties. *Vision research*, 94, 62-75.

Franzon, F., Zanini, C., & Rugani, R. (Forthcoming). Do non-verbal number systems shape grammar? Numerical cognition and Number morphology compared. To appear in *Mind and Language*.

Huntley-Fenner, G., Carey, S., & Solimando, A. (2002). Objects are individuals but stuff doesn't count: Perceived rigidity and cohesiveness influence infants' representations of small groups of discrete entities. *Cognition, 85(3)*, 203-221.

Jackendoff, R. (1991). Parts and boundaries. *Cognition, 41*, 9-45.

Katz, G. & Zamparelli, R. (2012). Quantifying Count/Mass Elasticity. Choi, J. et al. (eds). *Proceedings of the 29th West Coast Conference on Formal Linguistics*. Somerville, MA: Cascadilla Proceedings Project, 371-379.

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

Kulkarni, R., Rothstein, S., & Treves, A. (2013). A Statistical Investigation into the Cross-Linguistic Distribution of Mass and Count Nouns: Morphosyntactic and Semantic Perspectives. *Biolinguistics* 7, 132-168.

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

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

Mottaghi, R., Chen, X., Liu, X., Cho,N.-G., Lee, S.-W., Fidler,S., Urtasun,R., & 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*, 891–898.

Pelletier, F. J. (2012). Lexical Nouns are Neither Mass nor Count, but they are Both Mass and Count. D. Massam (ed.). *A Cross-Linguistic Exploration of the Count-Mass Distinction.* Oxford: OUP, 9-26.

Prasada, S., Ferenz, K., & Haskell, T. (2002). Conceiving of entities as objects and as stuff. *Cognition 83(2)*, 141-165.

Rothstein, S. (2010). Counting and the Mass/Count Distinction. *Journal of Semantics* 27(3), 343-397

Russakovsky, O., et al. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision 115(3)*, 211–252.

Schiehlen, M., & Spranger, K. (2006). The Mass–Count Distinction: Acquisition and Disambiguation. *Proceedings of the 5th International Conference on Language Resources and Evaluation* (Vol. 277).

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

Strickland, B. (2016). Language Reflects "Core" Cognition: A New Theory About the Origin of Cross-Linguistic Regularities. *Cognitive science.* DOI: 10.1111/cogs.12332

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

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

Zanini, C., Benavides, S., Lorusso, D., &. Franzon, F. (2017). Mass is more. The conceiving of (un)countability and its encoding into language in five year-old-children. *Psychonomic Bulletin & Review 24*(4) 1330-1340. Doi: 10.3758/s13423-016-1187-2