What's in a name? A large-scale computational study on how competition between names affects naming variation

Eleonora Gualdoni[a,*], Thomas Brochhagen[a], Andreas Mädebach[a], Gemma Boleda[a,b]

[a] *Universitat Pompeu Fabra, Roc Boronat 138, Barcelona, 08018, Spain*
[b] *Catalan Institution for Research and Advanced Studies (ICREA), Passeig Lluis Companys 23, Barcelona, 08010, Spain*

Abstract

Speakers often use different names to refer to the same entity (e.g., "woman" vs."tennis player"). We study how typicality affects variation in naming visually presented objects. We use a novel computational approach to estimate visual typicality from images, and analyze a large dataset containing naming data for realistic images. In contrast to previous work, we take into account the visual properties of both the object and the scene in which it appears; and factor in multiple candidate names. We show that visual typicality mediates competition between candidate names: high competition, induced by the relationship between the visual properties of the object and the visual representations associated to names, predicts higher naming variation. On a methodological level, we demonstrate the potential of using large-scale datasets with realistic images in conjunction with computational methods to shed light on how people name objects.

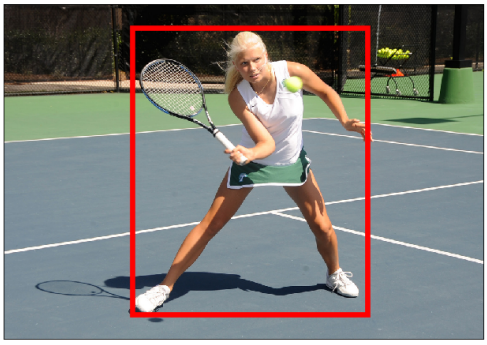
Keywords: object naming, naming variation, visual typicality, object typicality, context typicality, computational method.

*Corresponding author
Email address: eleonora.gualdoni@upf.edu (Eleonora Gualdoni)

1. Introduction

We refer to objects in most interactions. In doing so, we usually choose a word in our lexicon to name them, such as "woman" or "tennis player" for the persons in Figure 1. This involves cognitive processes that link the properties of the object with the lexicon. The mapping of an object's properties to the lexicon is, however, not one-to-one: different names can be used for the same object. In this article, we examine how the visual properties of objects and the contexts they appear in influence how varied speakers' choices are when naming them. In particular, we focus on the role of **visual typicality**, and on how different name alternatives **compete** as a function of typicality. In contrast to most previous studies, we analyze data from realistic images (i.e., real-world objects seen in meaningful contexts like those of Figure 1) and use computational methods adapted from the field of Computer Vision to estimate visual typicality. Our work is thus part of a very recent line of research that proposes using state-of-the-art Computer Vision representations to address questions about human language (Ahn et al., 2021; Guenther et al., 2021).

Naming variation has so far received relatively little attention in the literature. Naming norms, encompassing hundreds of objects, are available for a number of languages (e.g., Alario & Ferrand, 1999; Brodeur et al., 2010; Snodgrass & Vanderwart, 1980; Tsaparina-Guillemard et al., 2011). These norms were collected by asking subjects to freely produce a name for visually presented single objects in standardized image sets –i.e., in descriptive settings. In naming norms, the image sets are usually preselected with the goal to minimize naming variation by choosing easily identifiable and stylized depictions of a given object category (see Figures 2a and 2b for examples). Indeed, naming variation is often regarded as a nuisance variable in need of control for a given experimental task, not as a variable of interest (Alario & Ferrand, 1999; Brodeur et al., 2010, 2014). However, even in this scenario of prototypical objects presented in isolation, pervasive variation is still attested in subject responses. In our study, we look at naming variation as the variable of interest.



(a)

NAMES: **woman (17)**, tennis player (8), player (4), athlete (2).

VARIATION (H): 1.62

(b)

NAMES: **woman (30)**, tennis player (3), girl (2).

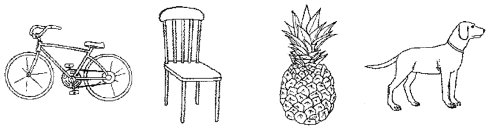
VARIATION (H): 0.73

Figure 1: Examples of images with top name "woman" and alternative name "tennis player" in ManyNames (Silberer et al., 2020a) (in parentheses, response counts; in bold face, the most frequent name, or *top name*). Image 1a exhibits more naming variation, expressed by the information statistic H (see Section 4).

There is also a sizable literature focused on discriminative tasks, with the standard paradigm consisting of an artificial scene constituted by different objects (Graf et al., 2016; Jescheniak et al., 2005, see Figure 2c for an example). The task of subjects is to produce a linguistic expression that uniquely identifies a target object. The focus of this line of research has been on how expressions for the same object vary as a function of which other objects are present in the scene (Graf et al., 2016; Jescheniak et al., 2005); or across interactions between interlocutors (Brennan & Clark, 1996; Haber et al., 2019). We instead analyze inter-speaker variation for the same object; and we do so in a descriptive setting akin to that of naming norms.

Lexical choices in object naming are tightly linked to the way humans categorize and conceptualize objects. Different speakers may arrive at different conceptualizations, resulting in naming variation. Early seminal studies on object categorization and naming (e.g., Jolicoeur et al., 1984; Rosch & Mervis, 1975; Rosch et al., 1976) concentrate on the level of specificity chosen in a taxonomy of categories; for instance, the choice between "animal", "dog", and "Dalmatian" for a given dog. This early work has been enormously influential, and subsequent research in Cognitive Science, both on categorization and on naming, has overwhelmingly focused on taxonomic aspects (Graf et al., 2016; Jescheniak et al., 2005). However, the names that speakers give to objects reflect not only their preferred taxonomic level, but also how they conceptualize them more broadly. This includes different conceptualizations of the same object (e.g. "woman" vs. "tennis player" in Figure 1; Ross & Murphy, 1999) and even disagreements as to which category an object belongs to in the first place (e.g. "woman" vs. "man" for the same person seen from afar; Silberer et al., 2020a). Our use of a large-scale dataset of realistic images enables us to encompass different sources of naming variation.

The body of research discussed above highlighted the role of typicality in categorization and naming. The term "typicality" in this case is usually applied to concepts (or categories). For example, ducks, in general, are atypical birds. However, typicality has also been shown to be relevant for instances



(a) Stimuli by Snodgrass and Vanderwart, 1980

(b) Stimuli by Brodeur et al., 2014

(c) Stimuli by Graf et al., 2016

Figure 2: Examples of stimuli employed in naming studies. In panel (a), the stylized black-and-white stimuli used by Snodgrass & Vanderwart (1980). In panel (b), the colored images used by Brodeur et al. (2014): Objects are more realistic but without any context. In panel (c), the stimuli used in the study by Graf et al. (2016): Here simple colored stimuli are arranged in grids to artificially generate a context for the object to name, highlighted in green.

(e.g., different images of ducks). Recall from above that naming norms are based on visually depicted objects –that is, specific graphical instantiations of a given concept, even if they are designed to be prototypical. Snodgrass & Vanderwart (1980) reported a significant correlation between naming variation and subjective ratings of image agreement, with the latter being defined as the resemblance between an experimental item and the mental image for this type of object. Thus, image agreement is arguably typicality applied to instances. In this paper, we use "visual typicality" instead of the more obscure "image agreement", to highlight the connection to typicality more generally.

In Snodgrass & Vanderwart's work, image agreement correlated negatively with naming variation: the less typical an object was for the target category, the higher the observed variation in subject responses. This result has been replicated in several subsequent studies (Alario et al., 2004; Brodeur et al., 2010; Shao & Stiegert, 2016; Liu et al., 2011; Tsaparina et al., 2011).[1] However, there are two important limitations when relating these findings to human naming variation in realistic scenarios, besides the aforementioned fact that they consist of highly idealized stimuli. The first is the fact that naming norms typically include only one instance for each category, which does not allow for an analysis of intra-category variation. The second is that typicality ratings have so far been collected only for the top name –the name most frequently produced by subjects–, and thus most analyses ignored other produced names.[2] An important reason for excluding less frequent names is that gathering subjective ratings for multiple object names per image is costly. Another –and related– reason is that norming data were usually collected and analyzed with the final goal of modelling latencies of *names* in object naming, e.g. "chair" or "apple"

[1] Of note, other variables have also been repeatedly shown to correlate with naming variation in at least some studies, most notably familiarity and age-of-aquisition (Alario et al., 2004; Liu et al., 2011; Moreno-Martínez & Montoro, 2012; Tsaparina et al., 2011).

[2] Koranda et al. (2018) and Vitkovitch & Tyrrell (1995) do analyze multiple alternative names, although not for typicality, but to assess how lexical choice and naming latencies, respectively, are affected by the availability of multiple candidate object names.

(e.g., Alario et al., 2004; Barry et al., 1997; Shao & Stiegert, 2016). Ratings for objects' candidate names beyond the top name were only of limited relevance for this goal. However, this exclusion is not conducive to explaining naming variation.

Finally, while previous work has focused only on the typicality of the object, we also analyze the typicality of the context in which the object appears. We define *context* as the scene the object is in (e.g., the tennis court in Figure 1a), which requires working with objects in realistic images.

To sum up, the present study investigates naming variation as a phenomenon in its own right. We aim at better characterizing variation in object naming by analysing multiple, varied, and realistic images for a given object category (as opposed to a single, stylized depiction), as well as multiple candidate names competing for lexical choice. Large part of the variation in our data results from inter-individual differences in object identification and conceptualization (e.g., deciding whether to highlight gender or action-related information when naming people, or which taxonomic level to use; Silberer et al., 2020a). We hypothesize that visual typicality impacts the degree of competition between candidate names, which is reflected in the amount of variation in speakers' naming choices. For instance, being fairly typical for the names "woman", "tennis player", and "athlete", like the person in Figure 1a, can trigger competition between candidate names, resulting in higher naming variation than in the case where the person is less typical for "tennis player" and "athlete" (Figure 1b). We expect similar effects for object and context typicality, although they may be less pronounced for the latter because contexts are likely less informative for a given name than the object itself.

2. Data availability statement

The original ManyNames data are available at `https://github.com/amore-upf/manynames`. Data and scripts for our analyses are available at `https://osf.io/s7h9f/?view_only=beaeaa70102b42a09f0a547377d5b320`.

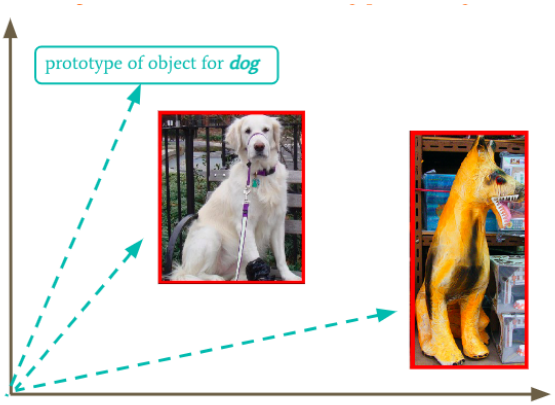


Figure 3: Computational method to estimate the typicality of a given image for a given name. High-dimensional vector representations of objects (in the figure, 2-D, for illustration) are averaged to obtain a visual prototype for a name. Typicality for a given image is the cosine similarity between the representation of the image and the name prototype. In this case, the golden retriever is closer to the prototype, so it is deemed more typical for the name "dog" than the toy dog on the right. The procedure to obtain typicality estimates for contexts is similar – see text.

3. Typicality estimation

As mentioned in the introduction, subjective ratings of visual typicality in naming norms are based on the similarity of the mental image evoked by a name and the image being evaluated. We use a computational equivalent of this procedure, also independently proposed by Guenther et al. (2021). The procedure is summarized in Figure 3 and detailed next.

3.1. Data: ManyNames

We use the ManyNames dataset (Silberer et al., 2020a), a large-scale resource containing up to 36 naming annotations for 25K objects in real-world images. We distinguish between the top name (the name most frequently used by the subjects) and the remaining, alternative names. The images in ManyNames were selected from VisualGenome (Krishna et al., 2017), a collection of 100K images often used for Computer Vision tasks. VisualGenome contains (single) name annotations for object. Naming annotations in ManyNames were collected

by asking human participants to freely produce a name to describe objects outlined by red bounding boxes, as illustrated in Figure 1.[3] To avoid sparsity in our computational representations, we only considered objects for which at least 20 naming annotations are available, resulting in 24.5K images.

3.2. Object typicality

Our computational estimate of the typicality of a given object for a given name is based on the distance between a visual representation of the object and a visual prototype for the name. We define the visual prototype of a name as the average visual features of images for which this name has been used, where the visual features are extracted with a Computer Vision model (see below). This operationalization follows the assumption that the prototypical exemplar of a category is the mental image of an average member of all the class exemplars (Gärdenfors & Williams, 2001; Rosch et al., 1976). This idea is consistent with image agreement as defined in naming norms, and in line with what has been done in recent computational studies on related phenomena (Ahn et al., 2021; Guenther et al., 2021; Xu et al., 2021).

Of the 1618 distinct object names in ManyNames, we built prototypes for the 874 that have at least 30 instances in VisualGenome (after excluding instances that also appear in ManyNames, to avoid circularity). Note, that, due to Zipf's law, the majority of excluded names are very rare (see Appendix A for their frequency distribution). Visual representations for individual objects were obtained using a state-of-the-art Computer Vision model trained on VisualGenome (Anderson et al., 2018). This model, like most current models in Computer Vision, is based on deep learning (LeCun et al., 2015), a type of model that learns representations for data as it learns to solve some task. The model we use was trained to perform two tasks: image captioning (providing

[3]Note that names that were used only once are not included in the dataset, due to constraints in the validation phase of the dataset (see Silberer et al. 2020b for details), except for those that are synonyms or hypernyms of the top name.

descriptions of pictures), and visual question-answering (answering questions about images). As part of carrying out these tasks, the model produces 2048-dimensional visual representations for the objects in an image. These are distributed representations, similar in nature to those for words in models such as Latent Semantic Analysis (Landauer et al., 1998) and distributional models more generally. As highlighted by Zhang et al. (2018), representations learnt by deep learning models trained to solve this kind of tasks correlate well with human perceptual similarity judgments, such that similar objects obtain similar representations.

The visual typicality of a ManyNames object for a name was then defined as the cosine similarity between the visual features of the object and the prototype for this name. Exemplifying the whole pipeline: To obtain the prototype for "tennis player", we (1) extracted all VisualGenome objects labeled "tennis player" (excluding images that are in ManyNames), where each object corresponds to a region in the image, such as the region marked in red in Figure 1a; (2) processed the objects with a Computer Vision model to obtain feature representations of them; (3) computed the prototype of "tennis player" by averaging all these feature representations. Then we (4) obtained estimates of typicality for individual instances by computing the cosine similarity between their feature representation and the visual prototype (with 0 being the lowest and 1 the highest value). For example, the object typicality scores obtained for Figures 1a and 1b for the name "tennis player" are, respectively, 0.77 and 0.67.

3.3. Context typicality

We obtained context typicality scores in an analogous fashion to the object typicality scores. The only difference is how we obtained a representation of the context. We aimed at a notion of context that synthesizes the global scene an object appears in, and adapted the procedure used by Anderson et al. (2018) for that purpose (see also Takmaz et al., 2022). Anderson et al. (2018) use their object detection module to detect 36 regions in an image, and average their visual features to obtain a representation of the whole scene to feed the image

Figure 4: Objects detected by Anderson et al. (2018) in an image from ManyNames. The red bounding box outlines the target object.

captioning model with. These regions include what one would commonly call an object (like a cat or a table), and also background elements like patches of grass or sky; see Figure 4 for example regions. We followed the same procedure, except that we excluded regions corresponding to the target object, since we wanted a representation of the context in which the object appears.

To exclude regions corresponding to the target object, we computed the intersection over union between the target and each of the 36 regions detected by the model. Intersection over union is the ratio between the overlapping area of two objects and their joint total area. This metric is commonly used in Computer Vision to evaluate object detection algorithms (Rezatofighi et al., 2019). We kept only regions with an intersection over union smaller than 0.1. We used the same metric to identify objects *without* context as well, i.e. objects that are almost as big as the entire image and for which a context typicality score would not be meaningful. We labeled all the objects whose bounding box had an intersection over union with the entire image higher than 0.77 as "objects

without context". This threshold was chosen through visual inspection: a value between 0.75 and 0.80 captures the majority of these cases. Except for the differences relating to the definition of context vs. target object, the pipeline for calculating context typicality is the same as that for object typicality. To exemplify its outcome, context typicality scores for Figures 1a and 1b for the name "tennis player" are, respectively, 0.82 and 0.43, aligning with our intuition for the typicality of the respective contexts.

3.4. Properties of the visual space

Figures 5 and 6 illustrate the relative positions of the object and the context name prototypes in a space defined by their visual features (reduced to 2D via Principal Component Analysis for plotting). The figures illustrate that the prototypes cluster meaningfully: Visual prototypes of semantically similar objects –and of contexts of semantically similar objects– tend to be similar to each other.[4] This in turn suggests that the visual feature space does relate to the mapping between the visual features of an object, or its context, and potential names for this object. Moreover, our computational estimates of object typicality seem to incorporate the typicality of the viewpoint, as the visual representations of objects with atypical viewpoint tend to be further from the prototype than the ones of objects with typical viewpoint –see Appendix B for example images. Some previous studies have teased apart the specific contribution of object viewpoint typicality from the role of object typicality in naming tasks (Brodeur et al., 2014; Johnson et al., 1996); future work should check how the two aspects can be differentiated computationally.

An exploration of typicality scores for the ManyNames objects additionally revealed systematic differences between object names from different levels of specificity, as could be expected from classic work on categorization (Rosch & Mervis, 1975; Rosch et al., 1976). This relation is illustrated in Figure 7.

[4] Also note that, according to the context space, humans, vehicles and buildings appear in the same contexts. This agrees with our intuitions.

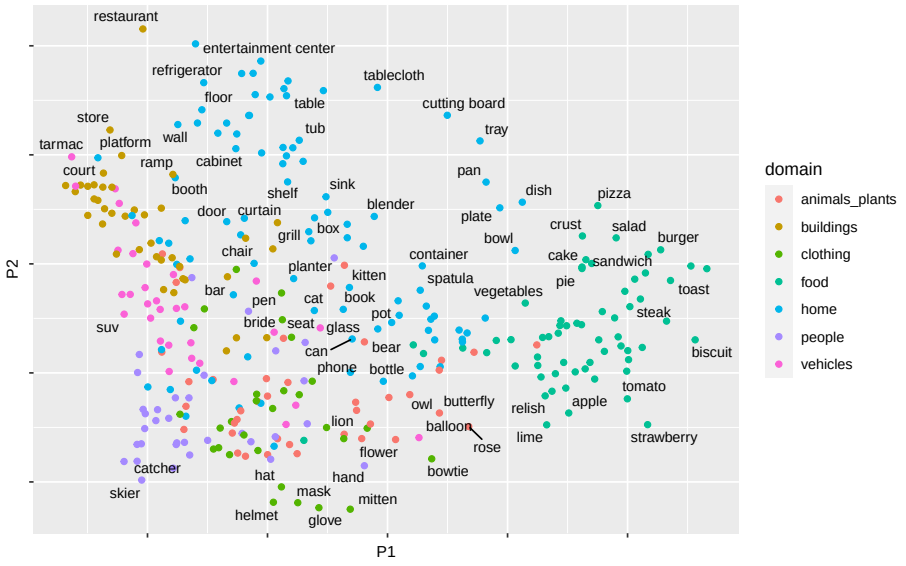


Figure 5: 2-D reduction of our space of object visual prototypes. For ease of visualization, only prototypes of top names are shown. Colors correspond to ManyNames domains.

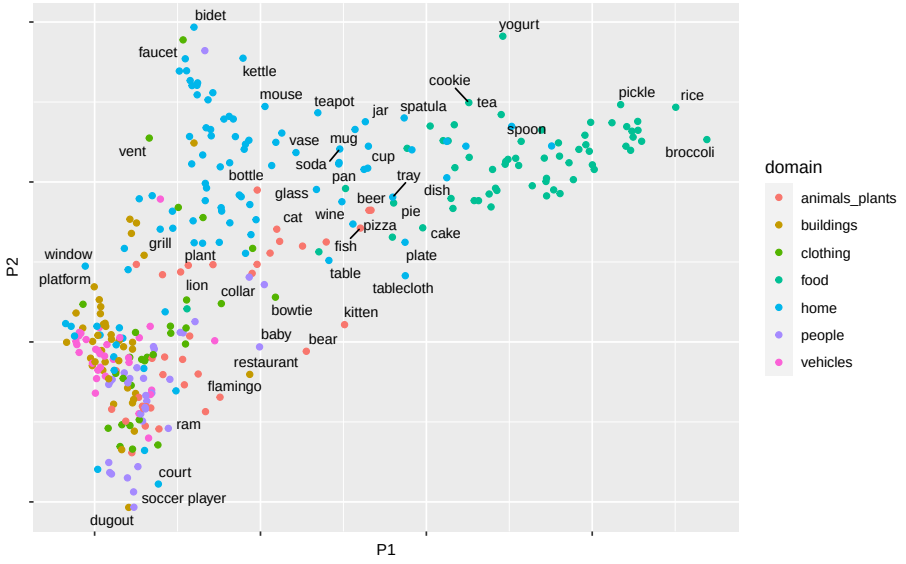


Figure 6: 2-D reduction of our space of context visual prototypes. For ease of visualization, only prototypes of top names are shown. Colors correspond to ManyNames domains.

For this illustration, we focus on the domain *people*, for which ManyNames contains many data points at different levels of specificity. The figure shows the cluster spread in both the object and the context visual spaces for the names in this domain. Cluster spread was defined as the average pairwise cosine distance between visual features of objects with the same top name –respectively, between visual features of contexts of objects with the same top name. We manually annotated the level of specificity of the names, ranging from the very general "person" (level 1) to more specific names, such as "umpire" or "catcher" (level 5). As shown in Figure 7, higher levels of specificity correspond to lower cluster spread –that is, objects with a specific name, like "skateboarder" or "catcher", are visually more similar to each other than objects with a less specific name like "woman" or "person". This mirrors the fact that more general names can be used for a more diverse set of objects than more specific names. Similarly, contexts for objects named "woman" or "person" are likely to be more diverse than the contexts in which skateboarders and catchers appear.

Taking stock, our exploration of the prototype space suggests that its geometry relates to properties of object names represented in this space. This suggests that the typicality scores and the relative position of prototypes may relate to the mapping between the visual features of an object and potential names for this object. This is further supported by the fact that the top name corresponded to the most similar object prototype for 24.1% of the objects in ManyNames (mean rank of the top name in terms of prototype similarity: 12.8 out of 874).

4. Analysis I: Naming variation and typicality

In this first analysis, we investigate how naming variation relates to the visual typicality of objects and their contexts. We restrict this analysis to the top name and the most frequent alternative name, e.g. "woman" and "tennis player" for the images in Figure 1. Analysis I seeks to test the computationally derived typicality scores by checking whether they replicate the result obtained

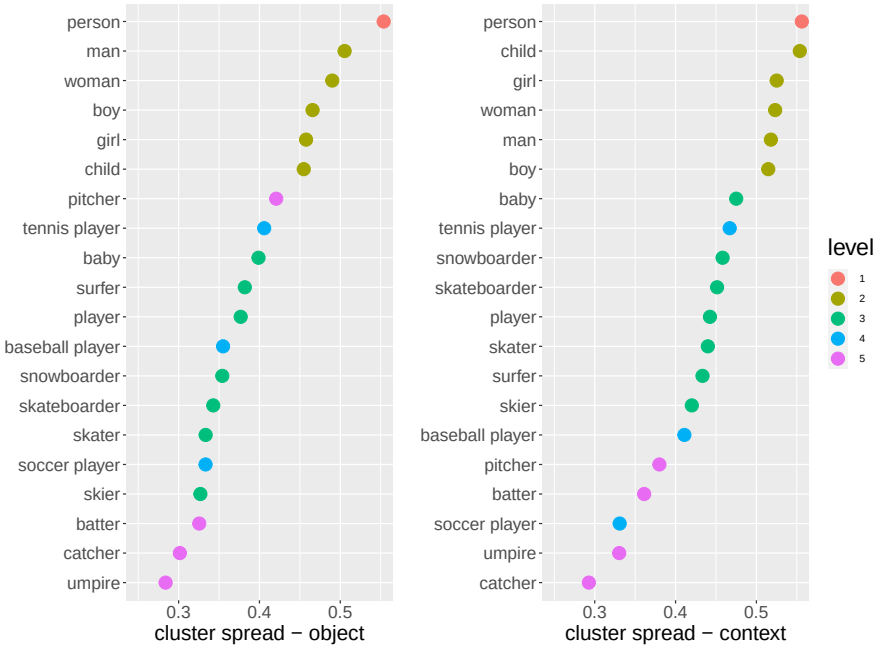


Figure 7: Cluster spread for names belonging to the domain *people*, computed as average pairwise distance between visual features of objects with that name as top name. Left: objects, right: contexts. Colors correspond to different levels of name specificity, from low ("person") to high ("umpire" / "catcher").

in prior work (Alario et al., 2004; Brodeur et al., 2010; Liu et al., 2011; Shao & Stiegert, 2016; Snodgrass & Vanderwart, 1980; Tsaparina et al., 2011) that lower variation is found for objects that are more typical of their top name. It also served as a first exploration of the role of alternative names and context typicality.

Our hypothesis was that the typicality of the alternative name would have an effect opposite to that of the top name. That is, we expected higher naming variation for objects that are more typical of their alternative name. This is because, assuming the same typicality for the most frequent name, relatively higher typicality for an alternative name should broaden the lexical options available for speakers, resulting in more inter-speaker variation. To illustrate, person (a) in Figure 1 is a more typical tennis player than person (b). Accordingly, more subjects use "tennis player" when naming her.[5]

With regard to context typicality, we expected to find the same pattern: A higher context typicality of the top name would have a negative effect on naming variation, and conversely for the alternative name. For instance, in Figure 1, the context of person (a) (a tennis court) is more typical for the alternative name "tennis player" than the context of person (b). This may contribute to the higher naming variation observed for person (a) compared to person (b).

Data and measures. We worked on the subset of ManyNames objects for which at least two different names were provided (17K out of 25K). Typicality estimates were derived as described above in Section 3. Naming variation was estimated in terms of entropy, as expressed by the information statistic H (Snodgrass & Vanderwart, 1980), defined as:

$$H = \sum_{i=1}^{k} p_i \; log_2(1/p_i), \tag{1}$$

[5]Notably, the top name for Figure 1a is still "woman". An effect of lexical frequency, as a proxy of lexical accessibility, may be at play in this case, as shown by Gualdoni et al. (2022). Moreover, Harrison (2022) reported a gender bias in ManyNames: annotators used sports-related names like "tennis player" much less for women than for men.

where k refers to the number of different names given to each object and p_i is the proportion of annotators giving each name. This measure captures information about the distribution of names across annotators. As exemplified in Figure 1, the person in panel (a) has a higher H score than that in panel (b) because she elicits more naming variation; both in terms of evoking more names and of having a more even spread of counts.

Regression model. We fitted a linear mixed-effects model with naming variation as the outcome variable and fixed effects for standardized object typicality and context typicality, each for both the top name and the alternative name. This made for four main effects in total. In the case of objects without context (see Section 3, 684 images in this analysis), we imputed context typicality with the average value in the data. Top names and alternative names were treated as random effects. By-topname and by-alternative name random slopes were included for object typicality and context typicality. Models were fit in R using *brms* (Bürkner, 2017; R Core Team, 2022), and diagnosed to rule out issues with our estimates. All diagnostics suggest a reliable model fit. Among others, all parameters have an $\hat{R} < 1.1$ (Gelman & Rubin, 1992); no saturated trajectories; and a large enough effective sample size (> 0.001 effective samples per transition).

4.1. Results

Fixed effect estimates are shown in Figure 8 and Table 1. Naming variation is higher the less typical an object is for its top name. Similarly, the object typicality estimate for the alternative name points towards direction we expected: higher variation the more typical an object is for its alternative name. However, this finding is not conclusive, since the confidence interval straddles 0. By contrast, when it comes to context typicality, counter to our expectations, we find no effect. This is true of both top and alternative names.

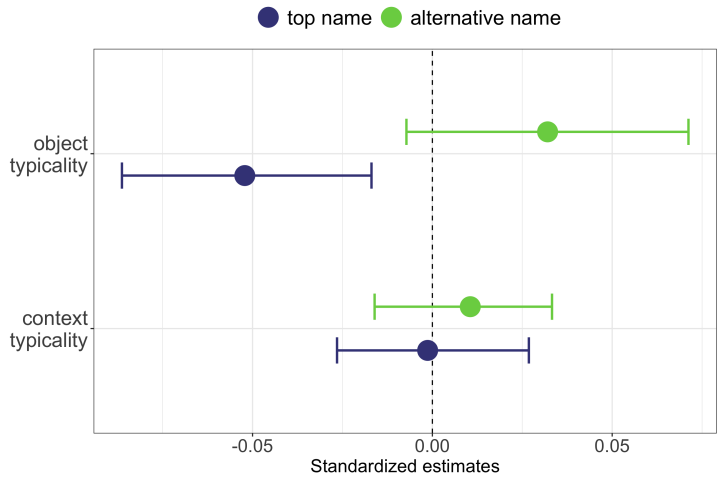


Figure 8: Fixed effect estimates. Bars correspond to 95% CIs. Positive vs. negative estimates show, respectively, the increase and decrease in naming variation for a one point difference in standard deviation of the predictor variable.

	Estimate	**Est.Error**	**l-95% CI**	**u-95% CI**
Intercept	1.33	0.03	1.27	1.39
Obj typ top name	-0.05	0.02	-0.09	-0.02
Obj typ alt name	0.03	0.02	-0.01	0.07
Ctx typ top name	-0.00	0.01	-0.03	0.03
Ctx typ alt name	0.01	0.01	-0.02	0.03

Table 1: Estimates of standardized fixed effects when predicting naming variation (H) as a function of object and context typicality.

4.2. Discussion

Analysis I replicates previous findings in showing a negative relationship between naming variation and object typicality (Brodeur et al., 2010, 2014; Liu et al., 2011; Moreno-Martínez & Montoro, 2012; Snodgrass & Vanderwart, 1980; Tsaparina-Guillemard et al., 2011). That is, people tend to choose the same name for an object when the object is very typical for that name. Importantly, we show this using a computational approach to estimate typicality. This suggests that our method offers a sensible and scalable way to address questions that, so far, had been approached with smaller data sets and more costly methodologies (i.e., human ratings).

One of the benefits of our computational approach is that it enabled us to investigate the way in which multiple candidate names jointly affect naming variation. Recall from the introduction that this aspect was neglected by previous studies that took into account the properties of only one name per object (Alario & Ferrand, 1999; Brodeur et al., 2010, 2014; Liu et al., 2011; Moreno-Martínez & Montoro, 2012; Snodgrass & Vanderwart, 1980; Tsaparina-Guillemard et al., 2011). Our results may be taken to suggest that the object typicality for an alternative name has, if any, the opposite effect than for the top name: The more typical an image is for an alternative name, the more likely it is that this name gets chosen over the top name. This would be in line with the idea that names compete for lexical selection. However, we want to emphasize that this finding is not certain; and we are restricting ourselves to considering only the most frequent alternative name. Since, in most cases, more than one name competes for lexical selection, considering the full range of competing names may yield a clearer picture. We address this issue in Analysis II.

Contrary to our expectations, Analysis I suggests that context typicality does not have an effect on naming variation in a descriptive object naming task such as the one the from ManyNames. The nature of the task may be key, since, when asked to freely produce a name for an object, speakers may not be influenced by the visual properties of the scene like they are in discrimination

tasks (e.g., Graf et al., 2016). Furthermore, prototypical contexts for different candidate names may also often be too similar as to affect name variation. For instance, the names "armchair" and "chair" are often naming alternatives for the same object, but the prototypical contexts for these two names are alike.

However, there is also the possibility that we merely fail to detect a true effect of context typicality due to how we represent contexts. The computational procedure we chose is robust in the sense that it has been shown to be a successful strategy to represent a scene for automatic image captioning and visual question answering tasks (Anderson et al., 2018). These tasks require a comprehensive representation of images. Additionally, the effectiveness of Anderson et al. (2018)'s model in extracting relevant visual features from images is supported by the fact that our results for object typicality replicate previous findings. However, due to the lack of previous research on context typicality, we cannot benchmark our computational estimates of context typicality using previous findings.

We thus turned to a comparison between subjective ratings of context typicality and our estimates. We collected human typicality judgments through crowdsourcing (see Appendix C for details), obtaining a positive correlation between our computationally-derived scores and average human typicality scores ($R = 0.43$). The average correlation between random pairs of human annotations is very similar ($R = 0.48$).[6] These results suggest that our computational context typicality scores are of good quality.

Finally, it could be that we fail to detect an effect of context typicality because of our limitation to two names per image –analogously to a potential explanation for the lack of a robust effect of alternative names. If the effect of context typicality is not very strong, its signal may not be picked up by this restrictive setup. We address this concern in Analysis II.

[6]Interestingly, for objects we find lower correlation between random pairs of human annotators (0.26), and an even lower correlation between our computational typicality scores and the average human scores ($R = 0.14$). See Appendix C for discussion.

5. Analysis II: Competition between names

Analysis I, while improving on previous work by including alternative names and context, is still based on a partial picture of naming behavior: on the one hand, for 39% of the objects in MN, speakers produced more than 2 different names; and, on the other, the analysis excluded the 31% of objects in Many-Names that received only one name.

In addition, Analysis I relied on knowing the names for a given image a priori. Therefore, model estimates from Analysis I cannot be used to predict naming variation without already knowing which names are used for the object. Finally, and importantly, the notions *top name* and *alternative name* are not static features of objects. They are themselves the result of competition.

In Analysis II we addressed these issues in the following way. Our general assumption is that the probability of selecting a given name for an image is a function of its visual similarity to the visual prototype of that name –that is, its typicality for the name. Naming variation is then expected to vary as a function of the number of viable (i.e., sufficiently typical) name candidates, with a larger number of viable candidate names leading to more naming variation due to higher competition. Figure 9 provides an illustration. Objects A and B are placed in different positions in the object visual space, based on their visual features. In particular, object A is much closer to the prototype "bench" than to any other candidate name. In contrast, object B is similarly close to 4 prototypes: "couch", "sofa", "chair", and "bench" (note that "couch" and "sofa" have almost overlapping prototypes, since they are synonyms). According to our model, specified below, this has consequences for the object names that speakers produce. For object A, the competition between names is dominated by "bench", as reflected in the object names showing no variation ($H = 0$). For object B, all 4 viable candidates are produced, resulting in higher naming variation ($H = 2.4$).

We implemented this idea using the prototypes in our visual space as potential attractors when naming an object. The position of the target object

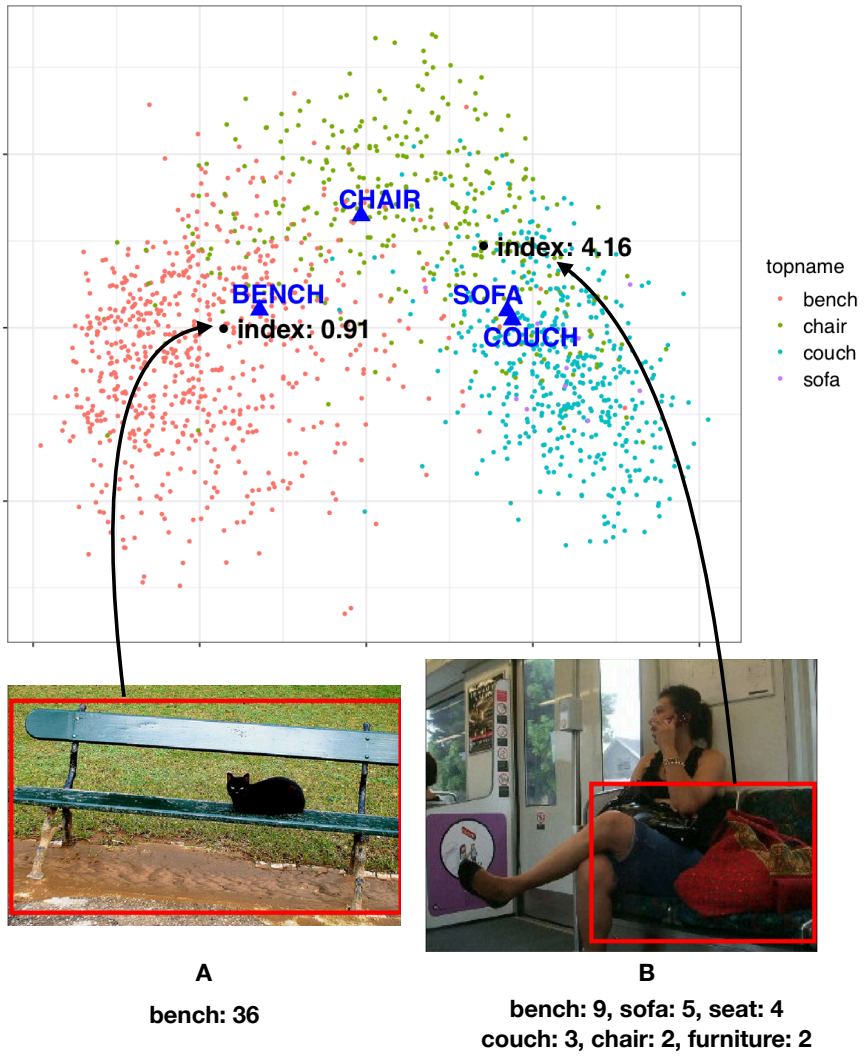


Figure 9: Visualization of the model proposed in Analysis II, obtained via a 2-D reduction of the region of our object visual space corresponding to images named "bench", "couch", "sofa", and "chair". Prototypes are represented by blue triangles. Dots represent objects. They are colored based on their top name. Images A and B show two ManyNames objects framed in a red bounding box. They are positioned in the visual space based on their visual features. Their index of object crowdedness is written in black. The names produced for the objects are listed below the images, followed by response counts.

relative to each of the prototypes in this visual space of attractors was then used to predict naming variation: The closer the image is to multiple visual prototypes for names, the more naming variation it is expected to evoke.

5.1. Methods

Data. We included the 24.5K ManyNames data points with more than 20 naming annotations available, to ensure robust estimates. We used the visual space populated by 874 prototypes and all the 24.5K individual images (see Section 3).

Index of crowdedness. We formalized competition via an *index of crowdedness*. This index quantifies, for each object image, the crowdedness of the area where it is located in our visual space. The crowdedness value for an image depends on how close the image is to each prototype in the visual space (see Figure 9). The closer more prototypes p are to a target image i, the higher the crowdedness for i. For instance, for object A in Figure 9 crowdedness is lower (0.91) than for object B (4.16). More concretely, the index is defined as:

$$crowdedness_i = \sum_p sim(i,p)^\gamma, \tag{2}$$

where i is an object in the data set, p is a prototype in our visual space, $sim(\cdot,\cdot)$ is cosine similarity, and $\gamma \geq 1$ is a temperature parameter. If $\gamma = 1$, then all prototypes contribute the same to the index, as a function of their distance to the image. As γ increases, closer prototypes exert more influence. This follows the intuition that competition may be non-linear in visual space ($\gamma > 1$), with competition between prototypes close to the target mattering more. Importantly, and by contrast to Analysis I, this index allowed us to treat all names as possible candidates of an object.

In analogy to Analysis I, we used both an *index of object crowdedness*, considering object visual features and object prototypes; and an *index of context crowdedness*, considering context visual features and context prototypes.

We expected the same effect for both object and context crowdedness: Objects/contexts in more crowded areas should elicit more naming variation. However, we expected a weaker effect in the case of contexts, since they may generally be more similar to each other than objects themselves, as discussed in Section 4.

Regression models. We identified the best γ-value for each index by fitting linear models with naming variation as outcome and the standardized index of object crowdedness as the sole predictor, for a sample of γ-values (1, 2, 5, 8, 10, 20, 30, 50). We selected the best γ-value through leave-one-out cross-validation, in terms of expected log predictive densities (Vehtari et al., 2017, 2019). Details and full model rankings for both object and context crowdedness are in Appendix D.

The best γ-value for the object index is 8, with values between 5 and 10 outperforming the rest by a large margin. This suggests a non-linear contribution of object prototypes to the competition between names. That is, prototypes near the target indeed contribute more than prototypes that are further away, corresponding to the intuition outlined above. Instead, for contexts the contribution of prototypes seems to be linear in visual space, with the highest ranked model being the one with $\gamma = 1$. More research is needed to elucidate whether this difference between visual prototype spaces is due to our particular operationalization, as discussed further below, or whether there is a true difference between them and the way they interact with naming choices. In what follows, we focus on the best models for each index, and we refer to the estimate derived from the best object/context γ-value simply as "index of object / context crowdedness".

We fitted a linear model with naming variation as outcome and both the index of object crowdedness and that of context crowdedness as predictors. This model outperforms single-predictor models with only object or only context crowdedness (see Table D.5 in Appendix D). As before, the model was diagnosed to rule out issues with the estimates. All diagnostics suggest reliable results.

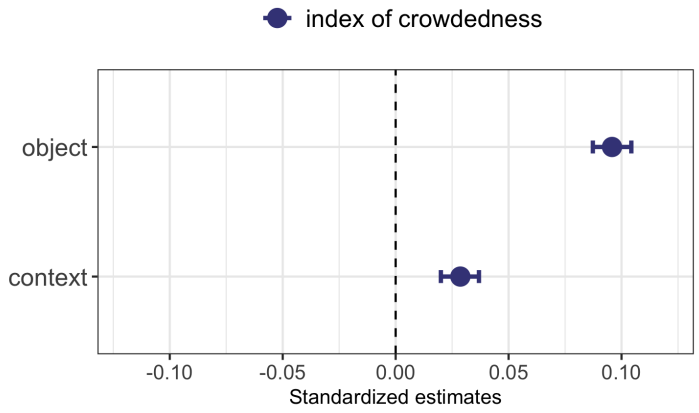


Figure 10: Effect estimates, showing the decrease or increase in naming variation for a one point difference in standard deviation of the predictor. Bars correspond to 95% CIs.

5.2. Results

Estimates are shown in Figure 10 and Table 2. Crowdedness affects variation in the way we expected: Naming variation is higher the more crowded the area is, in terms of name prototypes. Albeit to a lesser degree, the same is true of the object context: Context features that are close to many context prototypes elicit more naming variation. Taken together, this result suggests that, indeed, subjects tend to choose the same name for an object when there is less competition between naming alternatives; and that this competition is based on the visual properties of both the object and of the context in which it appears.

5.3. Discussion

Our second analysis expands the findings of the first one, considering all the candidate names in our lexicon as attraction points in a multidimensional visual space, and showing that naming variation increases with an increase in the competition of multiple candidate names. The approach in Analysis II has the additional advantage that there is no need to know the object names in advance to predict naming variation: The visual space has all the necessary information.

	Estimate	Est.Error	l-95% CI	u-95% CI
Intercept	0.71	0.004	0.70	0.71
Index of obj crowd	0.10	0.004	0.09	0.10
Index of ctx crowd	0.03	0.004	0.02	0.04

Table 2: Estimates of effects when predicting naming variation (H) using both indices of crowdedness.

In contrast to Analysis I, Analysis II suggests that the visual features of the context, and specifically how typical the context is for an object with a given name, do seem to affect naming variation after all (for a qualitative example, recall the case of the two women in Figure 1, where the tennis court is a more typical context for a tennis). We hypothesize that the structure of Analysis I, considering only the properties of the first and second most frequent names, may have led to too weak of a signal to detect the effect of context on naming. Indeed, context effects are much weaker than object effects in Analysis II. The same, at a smaller scale, may be at play when it comes to the effect of object typicality of alternative names: The fact that the index of object crowdedness is a good predictor of naming variation suggests that many naming alternatives compete for lexical selection, not just the top name and the most frequent alternative name.

The fact that the effect of context typicality is weaker than that of object typicality suggests that context prototypes are less informative than our object prototypes. This may be due to different reasons. First, as shown in Figure 11, the visual space of contexts is less spread out than the visual space of objects, possibly providing less discriminating information. Second, context prototypes of alternative names for the same object tend be more similar to each other (M=0.94, SD=0.05) than the corresponding object prototypes (M=0.81,

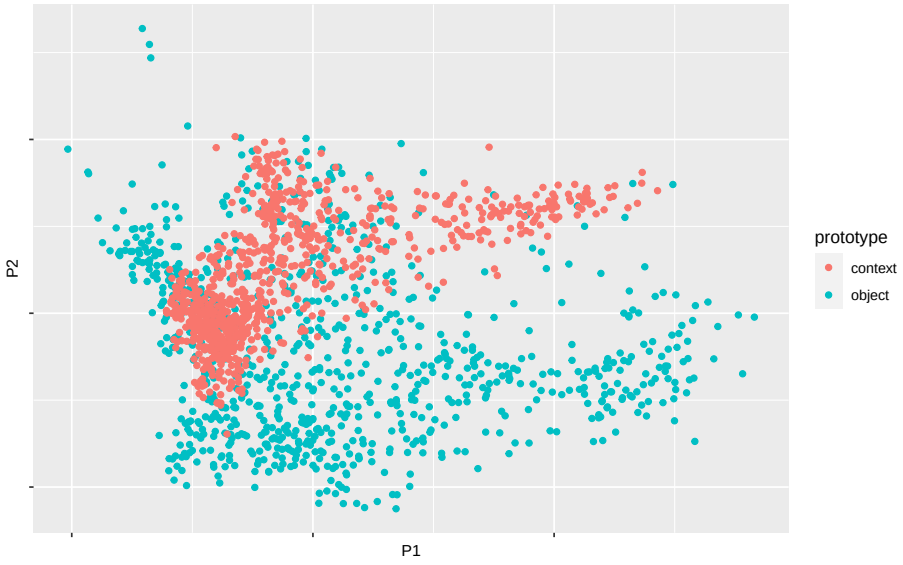


Figure 11: Visual object and context prototype spaces, after reduction to 2 dimensions.

SD=0.14), again pointing to the fact that they may carry less information. Figure 12 summarizes this trend through a side-by-side comparison of prototype similarities of both kinds. As discussed in Section 4, these properties could be an inherent property of contexts. These findings open promising avenues for future research.

Finally, recall that, with the index of object crowdedness, a non-linear version clearly characterizes the data better than a linear version. This confirms our intuition that closer prototypes should contribute more to the competition for a name than prototypes that are farther away. However, this prediction is not borne out in the case of context. At present, we have no hypothesis as to why this may be the case.

6. General Discussion

Objects can be called by many names. And yet, naming variation –inter-speaker variability in the names produced for a given object– has been either overlooked or considered as noise in most work in Cognitive Science (Alario &

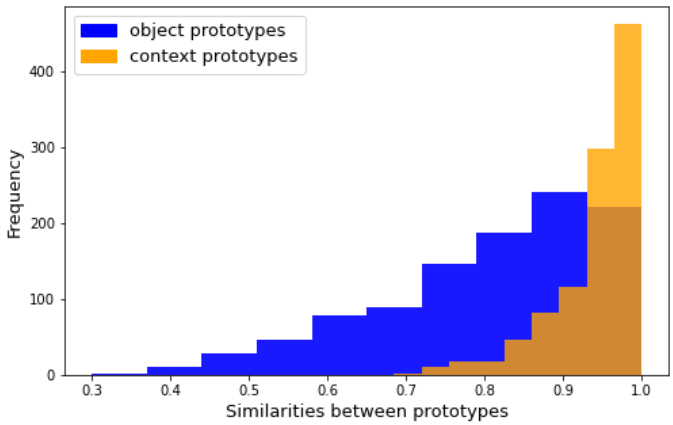


Figure 12: Histograms of prototypes' similarity between top name - alternative name pairs.

Ferrand, 1999; Brodeur et al., 2014; Tsaparina-Guillemard et al., 2011). Our work puts naming variation center stage, adding it to the repertoire of research questions in Cognitive Science. We believe that this is important because naming variation provides a window into how conceptual and linguistic knowledge interact in human behavior. Exploring naming variation thus promises to advance our knowledge of human naming behavior, and consequently also of language and cognition.

There are many different sources for naming variation. Some are conceptual, such as the different choices speakers face when categorizing an object ("woman" vs. "tennis player"), and others have more to do with the linguistic system, such as whether the lexicon offers (near-)synonyms to express a given concept ("sofa" and "couch"). We have tapped into one particularly relevant source of naming variation: How visually typical an object is for different names. Our hypothesis was that, because names compete for selection, when an object is visually typical for several names, higher variation would be observed. Conversely, lower variation would be observed when the object is typical for a single name (or simply fewer names). We tested this hypothesis using large-scale naming data for concrete objects in realistic images. The results support the hypothesis.

In particular, we replicate results from previous work about the role of typicality for a single name (Alario et al., 2004; Brodeur et al., 2010; Liu et al., 2011; Shao & Stiegert, 2016; Tsaparina et al., 2011), and add new insights on the role of typicality for multiple names, as well as on that of context typicality. We find that, in relation to the competition between names that they elicit, object and context typicality both have an effect on naming variation. The directionality of these effects is the same, with higher competition yielding higher variation, possibly for the same reasons. The effect of context, modelled as the scene in which an object appears, is however less strong, possibly because the visual context is less informative about the object than the visual features of the object itself.

We propose computational estimates of visual typicality as a less costly, time effective, and scalable alternative to the subjective ratings of visual typicality used in previous work on naming. More specifically, we have used data-driven distributed representations of images obtained through a Computer Vision model. Cognitive scientists have been successfully using data-driven distributed representations of words for decades, since the seminal work of Landauer & Dumais (1997) and Lund & Burgess (1996), among others. Classical word representations were generic lexical representations, and often used as surrogates for concepts[7] –i.e., they encode knowledge such as the fact that cats are more similar to dogs than to chairs. Computer Vision methods have developed over the last decade so as to be able to extract useful representations from images, that is, specific object *instances*: A given blond woman playing tennis in a tennis court, another one holding a racket against a grey background, etc. This affords new modeling possibilities, such as the ones explored here.

In particular, we used these representations to build both instance-like and concept-like visual representations: individual visual representations for object instances given particular names, and generic visual representations for names

[7]The relationship between word representations and concepts is nuanced; see Westera & Boleda 2019; Westera et al. 2021 for discussion.

by averaging individual representations, creating a prototype (Gärdenfors & Williams, 2001; Rosch et al., 1976). In this way, we could explore the interplay between instance-specific and concept-general aspects of naming, yielding a rich picture about the relationship between instances and concepts. This is key when conducting research on naming, since it is often the case that instances do not neatly fall into only one category, and the relationship between names and categories is not straightforward (Malt et al., 1999).

Our operationalization of prototypes and typicality, which is also found in recent computational approaches that leverage the information embedded in vector spaces (Ahn et al., 2021; Guenther et al., 2021; Xu et al., 2021), is however only one of many possible operationalizations. We do not intend it to be taken as a proposal reflecting the internal human processes at play when naming. Rather, its usefulness lies in its general motivations and its predictive acumen.

The results obtained with this method show that computational estimates of typicality are informative and can be used to predict naming variation. However, it is unclear to which degree human-rated typicality and computational estimates deviate systematically from each other. Comparing human ratings and computational estimates for a subset of our data shows clear correlations between the two measures (most pronounced for context typicality), but also substantial disagreement between them –see Appendix C for details. This may reflect that humans weight visual information to some extent differently than Computer Vision models (in agreement with findings from previous literature –for the case of shape bias, see Malhotra et al., 2020, 2022). It is conceivable that these differences arise in typicality judgments as well, with, for instance, humans relying less on size or view-point dependent visual information than machines. Moreover, as detailed in Appendix C, semantic knowledge about objects seems to contribute to human typicality judgments, adding another layer to the differences between human and computational typicality estimates. This interweaving of similarities and differences between how humans and machines represent visual information –and judge new stimuli based on the system of representations already stored– makes this topic a promising avenue for future

research. However, teasing apart where human-rated and computational estimates differ was beyond the scope of the present study.

7. Conclusion

In this work, we have presented a large-scale computational analysis of how people name objects, using naturalistic stimuli to investigate how visual typicality affects variation in naming. With respect to previous work, we have broadened the empirical coverage of our analysis along three axes: first, the amount of data (24.5K images and 874 names); second, the object names themselves, by including in the analysis all the names produced for a given object (as opposed to only the most frequent name); and third, the kinds of typicality explored, encompassing the typicality of both the objects and the scenes in which they occur. This increased coverage of factors affecting naming was achieved by adapting computational methods from Computer Vision to estimate visual typicality, instead of relying on subjective human ratings as done in previous work on naming.

We modelled visual prototypes as attraction points in a multi-dimensional space and found that naming variation can be predicted using the position of a given object in the visual space. Naming variation increases as a function of the density of suitable name candidates, with a larger number of similarly viable (i.e., similarly close) prototypes predicting higher naming variation. The same pattern was found for both object and context typicality, although less pronounced for context typicality.

Our results suggest that competition between candidate names is mediated by visual properties, and that this competition can be modelled using computational methods. Our approach provides a more flexible and less costly alternative to human rating data: It can be scaled to model large data sets and may facilitate research into aspects of human object naming that have not received sufficient attention to date, as we have done in the current paper for naming variation. More generally, our work is part of a budding strand of re-

search showcasing the potential of new Computer Vision methods for the study of human language (Ahn et al., 2021; Guenther et al., 2021).

Conflict of interest

The authors report no conflict of interest.

Acknowledgements

The authors thank the COLT research group for their useful feedback, as well as Carina Silberer for advice regarding Computer Vision models. 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 Ministerio de Ciencia e Innovación and the Agencia Estatal de Investigación (Spain; ref. PID2020-112602GB-I00/MICIN/AEI/10.13039/501100011033). This paper reflects the authors' view only, and the funding agencies are not responsible for any use that may be made of the information it contains.

References

Ahn, S., Zelinsky, G. J., & Lupyan, G. (2021). Use of superordinate labels yields more robust and human-like visual representations in convolutional neural networks. *Journal of Vision*, *21*, 13–13. URL: https://doi.org/10.1167/jov.21.13.13. doi:10.1167/jov.21.13.13.

Alario, F. X., & Ferrand, L. (1999). A set of 400 pictures standardized for french: Norms for name agreement, image agreement, familiarity, visual complexity, image variability, and age of acquisition. *Behavior Research Methods, Instruments, & Computers*, *31*, 531–552.

Alario, F. X., Ferrand, L., Laganaro, M., New, B., Frauenfelder, U. H., & Segui, J. (2004). Predictors of picture naming speed. *Behavior Research Methods, Instruments, & Computers*, *36*, 140–155. doi:10.3758/BF03195559.

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of CVPR*. arXiv:1707.07998.

Barry, C., Morrison, C. M., & Ellis, A. W. (1997). Naming the Snodgrass and Vanderwart Pictures: Effects of Age of Acquisition, Frequency, and Name Agreement. *The Quarterly Journal of Experimental Psychology Section A*, *50*, 560–585. doi:10.1080/783663595.

Brennan, S., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of experimental psychology. Learning, memory, and cognition*, *22 6*, 1482–93.

Brodeur, M., Dionne-Dostie, E., Montreuil, T., & Lepage, M. (2010). The bank of standardized stimuli (boss), a new set of 480 normative photos of objects to be used as visual stimuli in cognitive research. *PloS one*, *5*, e10773.

Brodeur, M., Guérard, K., & Bouras, M. (2014). Bank of standardized stimuli (boss) phase ii: 930 new normative photos. *PloS one*, *9*, e106953.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*, 1–28. doi:10.18637/jss.v080.i01.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*, 457–472. doi:10.1214/ss/1177011136.

Graf, C., Degen, J., Hawkins, R. X. D., & Goodman, N. D. (2016). Animal, dog, or dalmatian? level of abstraction in nominal referring expressions. *Cognitive Science*, .

Gualdoni, E., Brochhagen, T., Mädebach, A., & Boleda, G. (2022). Woman or tennis player? Visual typicality and lexical frequency affect variation in object naming. In J. Culbertson, A. Perfors, & V. Rabagliati, H. & Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society*. Cognitive Science Society. doi:10.31234/osf.io/34ckf.

Guenther, F., Marelli, M., Tureski, S., & Petilli, M. A. (2021). Vispa (vision spaces): A computer-vision-based representation system for individual images and concept prototypes, with large-scale evaluation, . doi:https://doi.org/10.31234/osf.io/n4dmq.

Gärdenfors, P., & Williams, M.-A. (2001). Reasoning about categories in conceptual spaces. In *Proceedings of the IJCAI* (pp. 385–392).

Haber, J., Baumgärtner, T., Takmaz, E., Gelderloos, L., Bruni, E., & Fernández, R. (2019). The photobook dataset: Building common ground through visually-grounded dialogue. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 1895–1910).

Harrison, S. (2022). Run like a girl: Sports-related gender bias in language and vision.

Jescheniak, J., Hantsch, A., & Schriefers, H. (2005). Context effects on lexical choice and lexical activation. *Journal of experimental psychology. Learning, memory, and cognition*, *31*, 905–20.

Johnson, C. J., Paivio, A., & Clark, J. M. (1996). Cognitive components of picture naming. *Psychological Bulletin*, *120*, 13–139.

Jolicoeur, P., Gluck, M. A., & Kosslyn, S. M. (1984). Pictures and names: Making the connection. *Cognitive Psychology*, *16*, 243–275.

Koranda, M., Zettersten, M., & MacDonald, M. C. (2018). Word frequency can affect what you choose to say. *Cognitive Science*, .

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D., Bernstein, M., & Li, F.-F. (2017). Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, *123*.

Landauer, T. K., & Dumais, S. T. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*, 211–240.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*, 259–284.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*, 436–44. doi:10.1038/nature14539.

Liu, Y., Hao, M., li, P., & Shu, H. (2011). Timed picture naming norms for mandarin chinese. *PloS one*, *6*, e16505. doi:10.1371/journal.pone.0016505.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic space from lexical co-occurence. *Behavior Research Methods Instruments & Computers*, *28*, 203–208. doi:10.3758/BF03204766.

Malhotra, G., Dujmović, M., & Bowers, J. S. (2022). Feature blindness: A challenge for understanding and modelling visual object recognition. *PLOS Computational Biology*, *18*, 1–27. URL: https://doi.org/10.1371/journal.pcbi.1009572. doi:10.1371/journal.pcbi.1009572.

Malhotra, G., Evans, B. D., & Bowers, J. S. (2020). Hiding a plane with a pixel: examining shape-bias in cnns and the benefit of building in biological constraints. *Vision Research*, *174*, 57–68. URL: https://www.sciencedirect.com/science/article/pii/S0042698920300742. doi:https://doi.org/10.1016/j.visres.2020.04.013.

Malt, B. C., Sloman, S. A., Gennari, S., Shi, M., & Wang, Y. (1999). Knowing versus naming: Similarity and the linguistic cat-

egorization of artifacts. *Journal of Memory and Language*, *40*, 230–262. URL: https://www.sciencedirect.com/science/article/pii/S0749596X98925931. doi:https://doi.org/10.1006/jmla.1998.2593.

Moreno-Martínez, F., & Montoro, P. (2012). An ecological alternative to snodgrass & vanderwart: 360 high quality colour images with norms for seven psycholinguistic variables. *PloS one*, *7*, e37527.

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*, 195–203. URL: https://doi.org/10.3758/s13428-018-01193-y. doi:10.3758/s13428-018-01193-y.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria. URL: https://www.R-project.org/.

Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., & Savarese, S. (2019). Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*, 573–605.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*, 382–439.

Ross, B., & Murphy, G. L. (1999). Food for thought: Cross-classification and category organization in a complex real-world domain. *Cognitive Psychology*, *38*, 495–553.

Shao, Z., & Stiegert, J. (2016). Predictors of photo naming: Dutch norms for 327 photos. *Behavior Research Methods*, *48*, 577–584. doi:10.3758/s13428-015-0613-0.

Silberer, C., Zarrieß, S., & Boleda, G. (2020a). Object naming in language and vision: A survey and a new dataset. In *Proceedings of the 12th Language Resources and Evaluation Conference* (pp. 5792–5801). Marseille, France: European Language Resources Association.

Silberer, C., Zarrieß, S., Westera, M., & Boleda, G. (2020b). Humans meet models on object naming: A new dataset and analysis. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 1893–1905). Barcelona, Spain (Online): International Committee on Computational Linguistics.

Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of experimental psychology. Human learning and memory*, *6 2*, 174–215.

Takmaz, E., Pezzelle, S., & Fernández, R. (2022). Less descriptive yet discriminative: Quantifying the properties of multimodal referring utterances via CLIP. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*.

Tsaparina, D., Bonin, P., & Méot, A. (2011). Russian norms for name agreement, image agreement for the colorized version of the Snodgrass and Vanderwart pictures and age of acquisition, conceptual familiarity, and imageability scores for modal object names. *Behavior Research Methods*, *43*, 1085–1099. doi:10.3758/s13428-011-0121-9.

Tsaparina-Guillemard, D., Bonin, P., & Méot, A. (2011). Russian norms for name agreement, image agreement for the colorized version of the snodgrass and vanderwart pictures and age of acquisition, conceptual familiarity, and imageability scores for modal object names. *Behavior research methods*, *43*, 1085–99. doi:10.3758/s13428-011-0121-9.

Vehtari, A., Gabry, J., Magnusson, M., Yao, Y., & Gelman, A. (2019). loo:

Efficient leave-one-out cross-validation and WAIC for Bayesian models. URL: https://mc-stan.org/loo r package version 2.2.0.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and computing*, *27*, 1413–1432.

Vitkovitch, M., & Tyrrell, L. (1995). Sources of disagreement in object naming. *The Quarterly Journal of Experimental Psychology Section A*, *48*, 822–848. doi:10.1080/14640749508401419.

Westera, M., & Boleda, G. (2019). Don't blame distributional semantics if it can't do entailment. In *Proceedings of the 13th International Conference on Computational Semantics - Long Papers* (pp. 120–133). Gothenburg, Sweden: Association for Computational Linguistics. URL: https://aclanthology.org/W19-0410. doi:10.18653/v1/W19-0410.

Westera, M., Gupta, A., Boleda, G., & Padó, S. (2021). Distributional models of category concepts based on names of category members. *Cognitive Science*, *45*. doi:10.1111/cogs.13029.

Xu, A., Stellar, J., & Xu, Y. (2021). Evolution of emotion semantics. *Cognition*, *217*, 104875. doi:10.1016/j.cognition.2021.104875.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. arXiv:1801.03924.

Appendix A. Histogram of occurrences of missing names

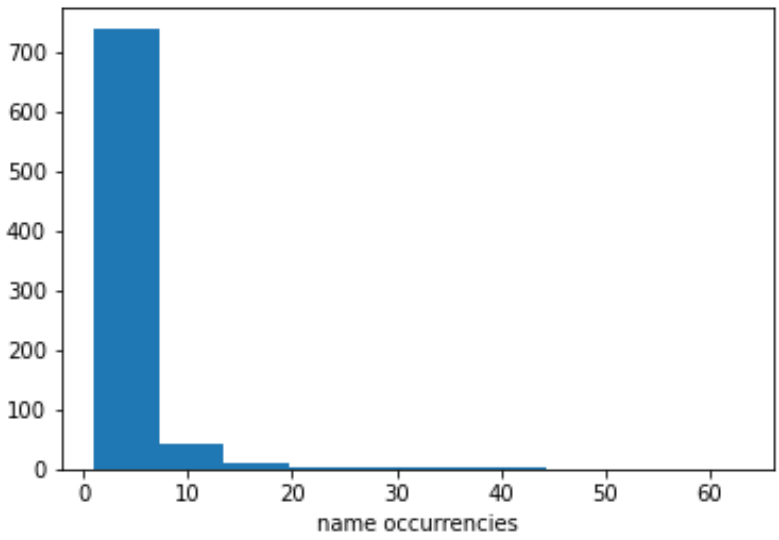
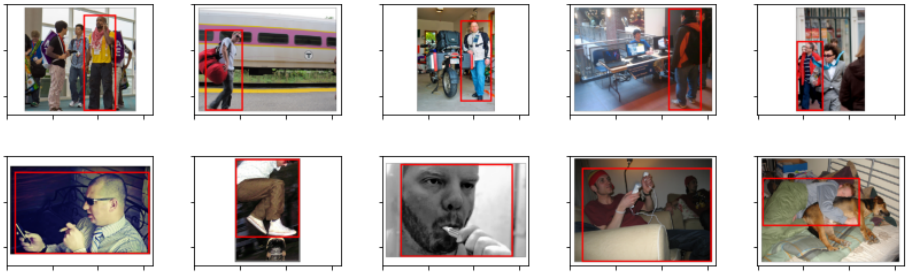


Figure A.13: Histogram of occurrences of names in ManyNames for which we lack prototypes.

Appendix B. Visual inspection of images with high / low computationally-derived typicality scores

Appendix B.1. Object typicality

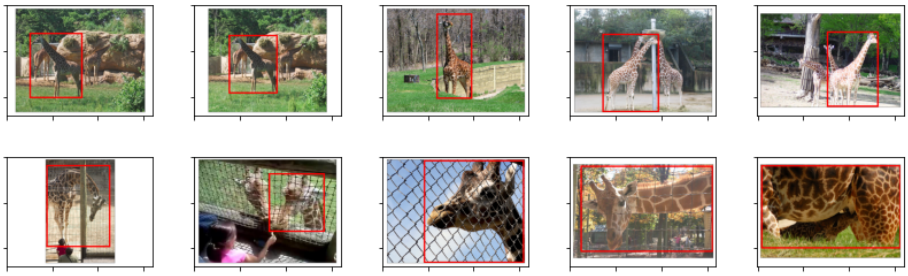
(a) woman



(b) man

(c) train

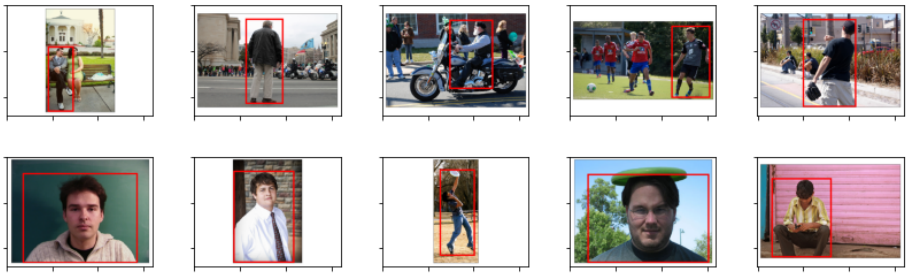
(d) shirt



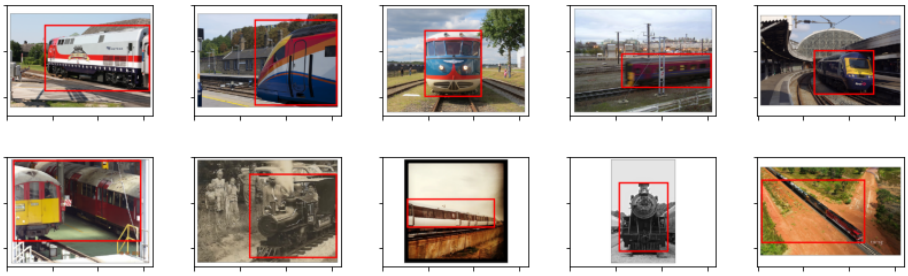
(e) giraffe

Figure B.13: The 5 most typical (first row of each panel) and the 5 most atypical objects (second row of each panel) based on our computationally-derived scores, for the 5 most frequently attested names in ManyNames. Typicality and atypicality grow going from left to right. (There are pairs of very similar images: of note, these are not repeated images. Small differences can be noticed displaying them in bigger dimensions –they are often different frames extracted from the same video.)

(a) woman

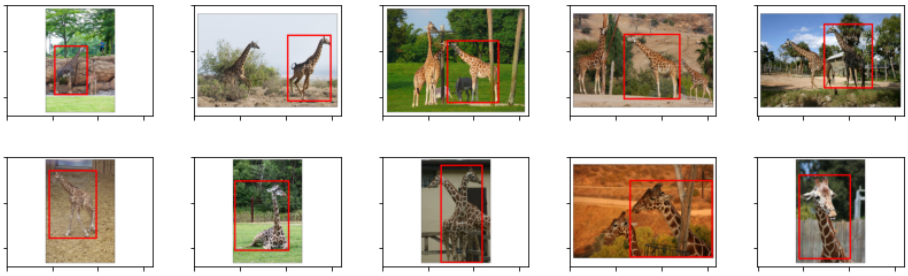


(b) man



(c) train

(d) shirt



(e) giraffe

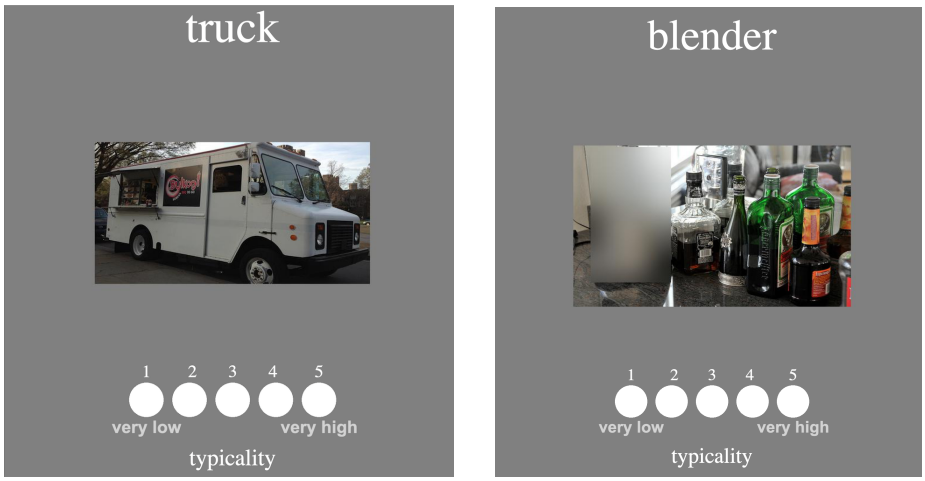
Figure B.13: The 5 most typical (first row of each panel) and the 5 most atypical context (second row of each panel) based on our computationally-derived scores, for the 5 most frequently attested names in ManyNames. Typicality and atypicality grow going from left to right. (There are pairs of very similar images: of note, these are not repeated images. Small differences can be noticed displaying them in bigger dimensions –they are often different frames extracted from the same video.)

Appendix C. Details on typicality judgment data collection

Design. We collected human judgments for object and context typicality to compare them with our computationally-derived typicality scores. We chose the subset to annotate as follows: For both objects and contexts, we selected 100 images by dividing the ManyNames images into 10 bins based on their computationally-derived typicality score (i.e. 10 percentiles). We sampled 10 images per bin, making sure that we did not repeat image top names; and avoiding to sample images whose top name is very frequent all from the same bin (to do so, we randomly sampled one image from the first bin, one image from the second, and so on).

Participants saw, in each screen, a single image with a name written above it. They were asked to give a typicality rating for it. Participants could give a score from 1 to 5 by clicking on radio buttons below the image, as illustrated in Figure C.14. In the object typicality task, subjects were presented with the cropped objects from ManyNames. They were instructed to give a score for how much the object looked like what they would expect when hearing or reading the corresponding object name. In the context typicality task, subjects were presented with the entire image from ManyNames; but with the object blurred so that they could focus primarily on the context. In the instructions, they were asked to give a score for how much the object surroundings looked like what they would expect to see for an object carrying the name written above. They could see one example before starting the task.

We collected 40 annotations for the context task, and 41 for the object task. Each participant annotated the entire set of 100 images, plus 5 randomly placed control items. The control items were chosen to monitor if annotators were paying attention to the task. They were intentionally picked to be very simple: We randomly sampled 5 object/context prototypes whose name did not appear in the list of top names previously selected as stimuli, and, for each of them, picked the furthermost object/context in terms of cosine distance (making sure we did not repeat items from the same category). This yielded, for instance,



(a) (b)

Figure C.14: Examples of screens shown to the participants. Panel (a) exemplifies the object typicality task. Panel (b) exemplifies the context typicality task.

the picture of a train, associated with the prototype "penguin"; the furthermost image from the prototype. Accordingly, these controls were expected to receive a very low score of typicality; serving as a benchmark of attentiveness. For the context control items, an additional step of manual selection was added to ensure their quality.

The data collection routine was written in Psychopy (Peirce et al., 2019) and launched through Pavlovia[8]. Participants were recruited via Amazon Mechanical Turk[9]. We only accepted annotators from the US, with HIT approval rate higher than 89% and number of approved HITs higher than 1000. We informed them that we would not collect any personal data (except for their workerID, that we would not make public), and that the goal of the experiment was to study how well certain names and images of everyday objects fit together. Moreover, they were informed that the task contained some control items designed

[8] `https://pavlovia.org/`

[9] `https://www.mturk.com/`

Figure C.15: Relationship between our computationally-derived typicality scores for objects (y-axis) and average human typicality scores (x-axis). The color gradient depicts the difference between the two ratings.

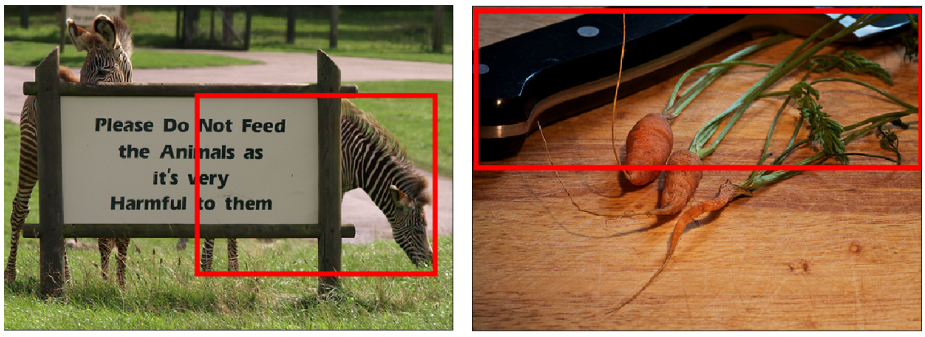
to ensure the quality of the annotation. Before being able to access the link to the experiment, participants had to complete an informed consent form. They were able to quit the experiment at any time. There was no time limit. We paid participants $3 for completing the task. The experiment was approved by the ethical board of Universitat Pompeu Fabra.

Results of the data collection. We excluded the data of participants that failed to give a low score (either 1 or 2) to more than one control, as this suggests that they were not paying enough attention to the task. This resulted in 41 out of 65 annotations being accepted for the object task; and 40 out of 84 annotations being accepted for the context task. For each item, we computed the average typicality score given by humans. Then, we computed the Pearson's correlation between the average human scores and the computational scores. For the object typicality scores, we obtained a correlation $R = 0.14$; this relationship is

Figure C.16: Relationship between our computationally-derived typicality scores for contexts (y-axis) and average human typicality scores (x-axis). The color gradient depicts the difference between the two ratings.

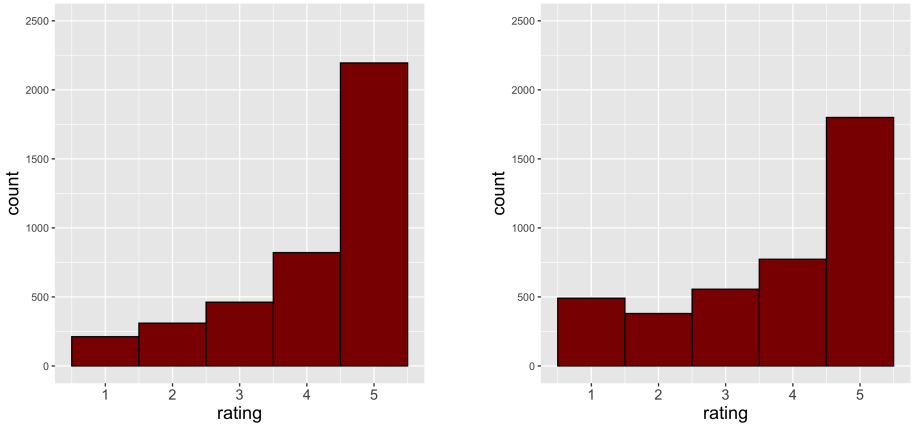
depicted in Figure C.15. For context typicality scores, we obtained a correlation $R = 0.43$; this relationship is depicted in Figure C.16.

As a baseline for comparison, we computed the correlation between random pairs of subjects, and averaged the resulting 20 correlations. When judging object typicality, the average correlation between randomly paired human annotators is 0.26; in the case of context typicality, this number grows to 0.48. These baselines put our results in perspective. They suggest that human participants did not agree much when judging the visual typicality of naturalistic images. Interestingly, this disagreement is higher when judging objects than contexts. In this light, the computationally-derived correlations of 0.14 for objects and 0.43 for contexts seem to be in line with the pattern shown by humans. The model agrees with the annotators to a similar degree than another human would do on the same task. Of note, human judgments about objects tend to be more skewed towards high typicality (see Figure C.17); and humans and models tend to agree more on what is *atypical* rather than on what is typical. After a manual inspection of the annotations, we noticed that humans seem to take into account some non-visual information as well, which models have no access to. For instance, a raw pizza is judged by humans as atypical. However, since it is visually similar to a cooked pizza, the computational judgment leans towards higher typicality. Moreover, humans appear to not factor in much viewpoint typicality. By contrast, our computational judgments clearly do: A dog seen through a net remains a fairly typical dog for humans, while it is judged atypical by the models –see Figure C.18 for other examples. While it makes sense to disentangle the judgment on object typicality from the judgment on viewpoint typicality for humans, the fact that our computational measure merges the two aspects is a desirable by-product in the context of our present analysis since viewpoint typicality has been shown to correlate with naming variation; and to do so in the same direction as object typicality (Brodeur et al., 2014; Johnson et al., 1996).



(a) (b)

Figure C.18: Examples of objects with atypical viewpoints. Panel (a): An object with top name "zebra"; with a human typicality score of 4.58 and a computational score of 0.61 (the average computational typicality for objects with top name "zebra" is 0.68). Panel (b): An object with top name "knife"; with human typicality score of 3.39 and a computational score of 0.38 (average computational typicality for objects with top name "knife" is 0.61).



(a) (b)

Figure C.17: Histograms of human ratings in the object typicality task - panel (a) - and in the context typicality task - panel (b)

Appendix D. Analysis II: model evaluations

We report here the model rankings based on leave-one-out cross-validation. Table D.3 and Table D.4 show, respectively, the rankings of models with index of object crowdedness and index of context crowdedness across γ parameters. Table D.5 shows the ranking of models with the best γ parameters, including the multi-variate model.

Appendix D.1. Index of object crowdedness: evaluation for best γ

model	ELPD difference	SE difference
idx obj 8	0.0	0.0
idx obj 5	-63.7	11.8
idx obj 10	-75.0	6.5
idx obj 2	-227.8	19.5
idx obj 1	-251.7	20.7
idx obj 30	-255.2	27.8
idx obj 50	-267.8	26.2
idx obj 20	-285.9	24.1

Table D.3: Ranking, based on leave-one-out cross-validation, of models fitted with index of object crowdedness with different values of γ (values of γ are made explicit in the model names). The second column shows differences in expected log-predictive densities to the highest ranked model; the third columns shows the standard error. The best model is boldfaced.

Appendix D.2. Index of context crowdedness: evaluation for best γ

model	ELPD difference	SE difference
idx ctx 1	0.0	0.0
idx ctx 2	-2.6	1.2
idx ctx 5	-12.5	3.7
idx ctx 8	-20.3	5.3
idx ctx 10	-24.0	6.1
idx ctx 20	-32.9	8.5
idx ctx 30	-38.5	9.7
idx ctx 50	-53.9	10.6

Table D.4: Ranking, based on leave-one-out cross-validation, of models fitted with index context of crowdedness with different values of γ (values of γ are made explicit in the model names). The second column shows differences in expected log-predictive densities to the highest ranked model; the third columns shows the standard error. The best model is boldfaced.

Appendix D.3. Index of crowdedness: evaluation for best model

model	ELPD difference	SE difference
idx obj + ctx	0.0	0.0
idx obj	-21.4	6.6
idx ctx	-249.0	21.9

Table D.5: Ranking, based on leave-one-out evaluation, of the best unifactorial models fitted with index of crowdedness and the multifactorial model (here referred to as "idx obj + ctx"). The second column shows differences in expected log-predictive densities to the highest ranked model; the third columns shows the standard error. The best model is boldfaced.