Woman or tennis player?
Visual typicality and lexical frequency affect variation in object naming

Eleonora Gualdoni (eleonora.gualdoni@upf.edu)
Universitat Pompeu Fabra, Barcelona, Spain

Thomas Brochhagen (thomas.brochhagen@upf.edu)
Universitat Pompeu Fabra, Barcelona, Spain

Andreas Mädebach (a.maedebach@gmail.com)
Universitat Pompeu Fabra, Barcelona, Spain

Gemma Boleda (gemma.boleda@upf.edu)
Universitat Pompeu Fabra / ICREA, Barcelona, Spain

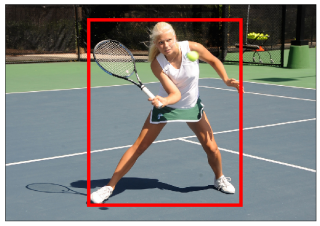
Abstract

Speakers often use different names to refer to the same entity (e.g., "woman" vs. "tennis player"). We here explore factors that affect naming variation for visually presented objects. We analyze a large dataset of object names with realistic images and focus on two factors: visual typicality (of both objects and the contexts they appear in) and name frequency. We develop a novel computational approach to estimate visual typicality, using image representations from Computer Vision models. Specifically, we compute visual typicality as similarity between the representation of an object/context to the average representation of other objects/contexts of its nominal class. In contrast to previous studies, we not only study the name used by most annotators for a given object (*top* name), but also the second most frequently used (*alternative* name). Our results show that the top name and the alternative name pull in opposite directions. People's naming choices are more varied for objects that are less typical for their top name, or more typical for their alternative name. They are also more varied when the top name has relatively low frequency (for alternative names, the opposite effect may be present but the data are not conclusive). Context typicality instead does not show a general effect in our analysis. Overall, our results show that visual and lexical characteristics relating to name candidates beyond the top name are informative for predicting variability in object naming. On a methodological level, we demonstrate the potential of using large scale datasets with realistic images in conjunction with computational methods to inform models of human object naming.

Keywords: object naming, naming variation, visual typicality, object typicality, context typicality, lexical frequency

Introduction

We successfully 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 people in Figure 1. This involves complex cognitive processing that allows us to link the properties of the object to our lexicon. The mapping of a representation of the object to the lexicon is not one-to-one: Often, different names can be used for the same object. In particular, recent work has shown that, while subjects do on average have a preferred name for objects, naming variation is pervasive (Silberer, Zarrieß, & Boleda, 2020). This variation corresponds not only to changes in conceptual taxonomic levels (e.g. "person" vs. "woman"), explored in early studies in psycholinguistics (Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976; Jolicoeur, Gluck, & Kosslyn, 1984), but also to different conceptualizations of the same object



(a) **woman (17)**, tennis player (8), player (4), athlete (2)
H: 1.62

(b) **woman (30)**, tennis player (3), girl (2).
H: 0.73

Figure 1: Examples of images with top name "woman" and alternative name "tennis player" in ManyNames (Silberer et al., 2020) (in parentheses, response counts). Image 1a shows more naming variation, expressed by the information statistic *H* (see *Methods*).

(e.g. "woman" vs. "tennis player" in Figure 1; Ross & Murphy, 1999); or even to disagreements as to what the object is (e.g. "woman" vs. "man" for the same person; Silberer et al., 2020). We here aim to better understand the factors that influence naming variation of visually presented objects.

Naming is a widely used task in different areas of cognitive science. It is of particular importance for studies on human language production (e.g., Humphreys, Riddoch, & Quinlan, 1988; Levelt, Schriefers, Vorberg, Meyer, & et al, 1991; Glaser, 1992). This kind of work focuses on the process that goes from a chosen name to its realization, e.g. in speech. Therefore, in this line of research, objects are often sought to elicit as little naming variation as possible. Norming studies providing standardized sets of images for these studies accordingly focus on stylized images of isolated objects, usually only with one object instance per category, designed to be a prototypical representation of the category (e.g., Snodgrass & Vanderwart, 1980; Brodeur, Dionne-Dostie, Montreuil, & Lepage, 2010; Brodeur, Guérard, & Bouras, 2014) – see Figures 2a and 2b. Moreover, in most of these studies naming

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

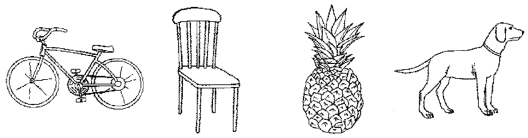
variation is regarded as noise; and only the name most frequently chosen for each object is subjected to analysis.

We instead look at naming variation as a phenomenon in its own right, and aim at characterizing how object **instances** are named, with all their idiosyncratic properties, as opposed to categories. We accordingly analyze naming data for objects in naturalistic images, exemplified in Figure 1. Moreover, in our analyses we take into account all the names produced, not only the most frequent name, so as to obtain a more comprehensive picture of not only the different naming possibilities for a given object, but also the factors that affect naming preferences for individual objects.

Context effects are an important aspect of many picture naming studies. It has been shown that context stimuli, such as distractor objects in a scene, affect lexical choices in referential tasks (e.g., Graf, Degen, Hawkins, & Goodman, 2016; Jescheniak, Hantsch, & Schriefers, 2005, see Figure 2c). For instance, this kind of study can contrast a scene with 3 different types of animals, one dalmatian, one greyhound, and one bear (see Figure 2c) to other scenes containing different types of objects. In these studies, the phenomenon of interest is the level of specificity of the name (e.g. "dalmatian" vs. "dog"), and the stimuli are explicitly designed to investigate this – with two clear options as to how to name a given object. Again, this research typically uses artificial scenes, with stylized objects placed side by side, and the objects are prototypical for their categories. We investigate a broader (in fact, open) set of possible naming choices, and a different notion of context, as explained below.

Specifically, we examine three factors: the **visual typicality**[1] of the **object**; the visual typicality of the **context** the object appears in; and the **lexical frequency** of the name, as a proxy of ease of lexical access (Alario & Ferrand, 1999; Koranda, Zettersten, & MacDonald, 2018).[2] We check how these factors relate to naming variation when taking into account not only the most frequent name, or *top name* ("woman" in the images in Figure 1), but also the most frequent alternative, or *alternative name* ("tennis player"). Alternative names have usually been neglected in previous work (for exceptions see Koranda et al. (2018); Vitkovitch and Tyrrell (1995)) and to the best of our knowledge have not been studied in terms of visual typicality, or in relation to naming variation.

We also examine for the first time the effect of the visual typicality of the context, defined as the global scene in which the object appears (e.g., the tennis court in Figure 1a, which is a very typical context for a tennis player). This is different from the role of distractor objects (examined in previous work, e.g. Graf et al., 2016), which we address only implicitly and indirectly, through their presence in the scene (see *Methods* for details).



(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.

Last but not least, we also introduce a methodological innovation. Previous work has used subjective human ratings to measure typicality. This is a costly and time-consuming procedure, which partially explains the focus of previous research on only one name per object. We instead rely on computational representations of the visual stimuli that allow us to automatically estimate typicality. This in turn allows us to expand the scope of inquiry to multiple names per image, and to the visual properties of both the objects and the contexts they appear in.

As for our expectations, in light of previous results we expect lower variation with increasing object typicality for the top name (Snodgrass & Vanderwart, 1980; Brodeur et al., 2010, 2014; Liu, Hao, li, & Shu, 2011; Moreno-Martínez & Montoro, 2012; Tsaparina-Guillemard et al., 2011). We furthermore hypothesize that the typicality of alternative names will have the opposite effect: The more typical an object is for the alternative name, the more competition can be expected to take place between the alternative name and the top name, thus increasing naming variation. This intuition is exemplified by a comparison of the two stimuli in Figure 1. The woman in Figure 1a is a more typical tennis player than

[1] Our notion of visual typicality is related to the notion of image agreement often reported in the literature (Snodgrass & Vanderwart, 1980; Tsaparina-Guillemard, Bonin, & Méot, 2011), but differs from it in some aspects. Image agreement is assessed through subjective comparisons of a presented image to a mental image evoked by a name. We instead assess the visual typicality of an object for a name through a comparison of its visual representation with the average object carrying that name.

[2] Of note, these are not the only aspects affecting naming variation: other factors not included in this analysis, such as familiarity or age of acquisition, play a role as well (for a review, see Johnson, Paivio, & Clark, 1996).

the woman in Figure 1b; and more subjects named the target object "tennis player" and fewer "woman" in Figure 1a compared to Figure 1b. For the same reason as for object typicality, we expect that the lexical frequency of the names will affect naming variation in opposite directions for top and alternative names, with lower naming variation when the top name is frequent, and an opposite effect when the alternative name is frequent, since that increases the competition with the top name.

Our analysis of the effect of context typicality on naming is more exploratory in nature. We tentatively extend our predictions for object typicality to visual context, expecting lower variation for objects in more typical visual contexts. In further analogy to object typicality, we also expect increased variation when the context typicality is higher for the alternative name – as suggested by Figure 1. We expect such contextual effects to be less pronounced than those of object typicality, given that contexts are likely less informative for a given name than the object to be named itself (see *Discussion*).

Methods

Data We analyzed data from the ManyNames dataset (Silberer et al., 2020), which provides up to 36 naming annotations for 25K objects in naturalistic images. These annotations were collected by asking subjects to freely produce a name to describe objects outlined by bounding boxes, as illustrated in Figure 1. We subset this data to the 17K objects for which at least two names are provided. We do so to model the potential competition between top and alternative names[3].

Typicality and frequency estimates Frequency estimates for names were extracted from SUBTLEX-US, a subtitle corpus of American English (Brysbaert & New, 2009).

As for object typicality, we first built visual prototypes for the names; then, to assess the typicality of object instances for their names, we computed their similarity to the prototype of those names, with the following procedure.

A prototype for a given name was defined as the average visual representation of the object images with that name. 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 (Rosch et al., 1976; Gärdenfors & Williams, 2001). Objects for a given name were selected from VisualGenome (Krishna et al., 2017), the resource from which ManyNames images were extracted. We excluded the objects that are also in ManyNames in the computation of prototypes, to avoid circularity. Also, to avoid noisy or biased prototype representations due to data sparsity, names for which VisualGenome provides less than 30 objects were excluded from the analysis. This excludes 770 out of 17K data points.

We create the visual representations for the objects using a state-of-the-art Computer Vision model trained on VisualGenome (Anderson et al., 2018). This model is trained to perform two tasks: image captioning (outputting a description of a picture), and visual question-answering (answering a question about the image). It incorporates a model that detects and labels objects in images (Ren, He, Girshick, & Sun, 2015). As part of carrying out the relevant tasks, thus, the model produces visual representations for the objects in the image. These are distributed representations, similar in nature to those for words in models such as LSA (Landauer, Foltz, & Laham, 1998) and distributional models more generally.

As estimate for object typicality, we used the cosine similarity between the object features and the prototype of its names. This enables us to track the effect of two visual typicality estimates on naming: one for the top name, and another for the alternative name.[4]

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

To exemplify the whole pipeline, for instance, 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 the Computer Vision model to obtain feature representations; (3) computed the prototype for "tennis player" by averaging all these feature representations; and (4) obtained estimates of typicality for individual instances by computing the cosine similarity between their feature representation (also created with the Computer Vision model) and the visual prototype. For example, the object typicality scores obtained for Figures 1a and 1b for the alternative name "tennis player" are, respectively, 0.77 and 0.67. The space of

[3]The data and the scripts used for the analysis are provided here: https://osf.io/q72ne/

[4]Typicality of the object viewpoint is often listed among the factors affecting naming tasks (Brodeur et al., 2014; Johnson et al., 1996). Our computational estimates of object typicality incorporate it: the visual representations produced by Anderson et al. (2018)'s computational model for objects with atypical viewpoints are more distant from the prototype than those produced for objects with typical viewpoints. This was confirmed by a qualitative inspection of the objects judged as very typical/atypical.

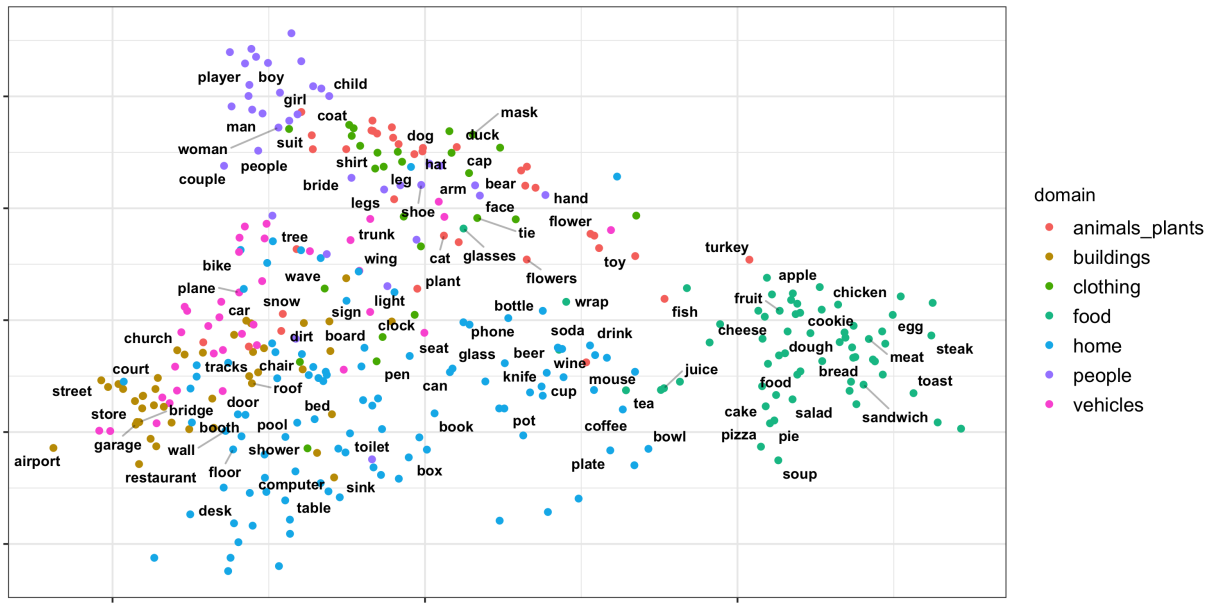


Figure 3: 2D reduction of our space of object visual prototypes (only top names are plotted for ease of visualization).

object visual prototypes thus obtained is shown in Figure 3, illustrating its relative cohesiveness in terms of domains.

We obtain context typicality scores in an analogous fashion; the only difference is how we obtain the representation of the context of an object instance. We aimed at a notion of context that synthesizes the *global scene*, and we reused a procedure used by Anderson et al. (2018) for that purpose (a similar procedure is also found in Takmaz, Pezzelle, and Fernández (2022)). Anderson et al. (2018) use the object detection module to detect 36 regions in the image, and average their representation to obtain a representation of the whole scene. 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. Anderson et al. (2018) follow this procedure to obtain a global representation of the image, which is then used by the image captioning model to produce a description. We follow the same procedure, except that we excluded regions corresponding to the target object, since we want a representation of the context in which it appears. We excluded the relevant regions by computing the intersection over union between the target object and each detection, that is, the ratio between the overlapping area of the objects and their total area, and keeping only regions with intersection over union smaller than 0.1. Intersection over union is commonly used in Computer Vision to evaluate the performance of object detection algorithms in identifying objects (Rezatofighi et al., 2019). To exemplify, the resulting context typicality scores for Figures 1a and 1b for the alternative name "tennis player" are, respectively, 0.82 and 0.43.

Naming variation estimates Naming variation for objects was estimated in terms of entropy, as expressed by the information statistic H (Snodgrass & Vanderwart, 1980), defined as:

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

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 object in image (a) is assigned a higher H score than the object in image (b), because it elicits more naming variation (in this case, both more names and 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, context typicality, and log-frequency, each for both the top name and the alternative name. Top names and alternative names were treated as random factors. By-topname random slopes were included for object typicality, context typicality, and alternative name frequency. By-alternative name random slopes were included for object typicality, context typicality, and topname frequency.[5]

[5]In brms/lme4 syntax (Bürkner, 2017): $H \sim$ obj typ top + obj typ alt + freq top + freq alt + ctx typ top + ctx typ alt + (1 + obj typ top + obj typ alt + freq alt + ctx typ top + ctx typ alt | topname) + (1 + obj typ top + obj typ alt + freq top + ctx typ top + ctx typ alt | altname)

	Estimate	Est.Error	l-95% CI	u-95% CI
Intercept	1.27	0.04	1.20	1.33
Obj typ top	-0.09	0.02	-0.12	-0.06
Obj typ alt	0.09	0.02	0.05	0.12
Ctx typ top	0.00	0.01	-0.02	0.03
Ctx typ alt	0.00	0.01	-0.02	0.03
Log freq top	-0.11	0.03	-0.18	-0.05
Log freq alt	0.02	0.02	-0.02	0.07

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

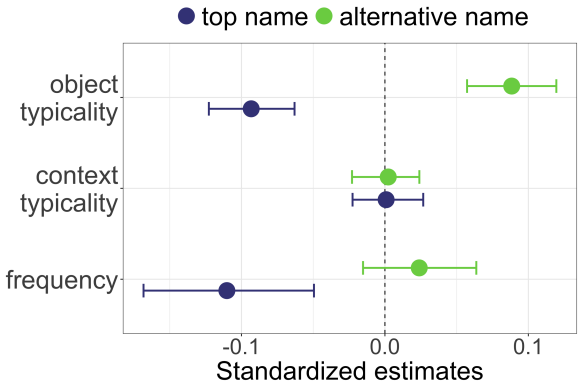


Figure 5: Fixed effect estimates. Bars correspond to 95% CIs. Positive vs. negative estimates show, respectively, the increase and decrease in naming variation for a one point difference in standard deviation of the predictor variable.

Results

Fixed effect estimates are shown in Figure 5 and Table 1. The model was also diagnosed to rule out issues with our estimates. All diagnostics suggest reliable results. Among others, all parameters have an $\hat{R} < 1.1$; no saturated trajectories, no divergent iterations; and a large enough effective sample size (> 0.001 effective samples per transition).

Object typicality for top name and alternative name affect variation in the way we expected: Naming variation is lower the more typical an object is for its top name, and higher the more typical it is for the alternative name. Also, as hypothesized, a more frequent top name is predictive of lower naming variation. When it comes to the effect of the frequency of alternative names, the trend we expected is not conclusively identified. Finally, counter to our expectations, we find no fixed effect for context typicality. This is true of both top and alternative names.

Taking stock, these results suggest that more people tend to choose the same name for an object when the object is very typical for that name, and if that name is very frequent. In contrast, naming variation increases the more typical the object is for an alternative name.

Discussion

Our large-scale computational analysis adds evidence to previous findings about object naming. We confirm prior knowledge about object typicality, showing that higher object typicality for top names is predictive of lower naming variation (Snodgrass & Vanderwart, 1980; Brodeur et al., 2010, 2014; Liu et al., 2011; Moreno-Martínez & Montoro, 2012; Tsaparina-Guillemard et al., 2011). Moreover, in line with Alario and Ferrand (1999), we find an effect of lexical frequency: Less naming variation is associated with objects whose top name has higher lexical frequency. The fact that we replicate results from past research suggests that our approach, using a computational approach deployable at large scale, offers an adequate and scalable way to address questions that, so far, had been approached with small data and more costly methodologies.

This computational approach enabled us to investigate two new factors: First, the way in which multiple candidate names jointly affect naming variation. In particular, our results show that the properties of alternative names have opposite effects with respect to the properties of top names: Higher object typicality for the top name relates to lower variation, whereas higher object typicality for the alternative name relates to higher variation. Similarly, higher top name frequency relates to lower variation, whereas higher alternative name frequency appears to yield, if anything, higher variation, but this is not conclusive in the present analysis. This is in line with the idea that names compete for lexical selection (for a review see Spalek, Damian, & Bölte, 2013). This aspect was neglected by previous studies that took into account the properties of only one name per object (Snodgrass & Vanderwart, 1980; Brodeur et al., 2010, 2014; Liu et al., 2011; Moreno-Martínez & Montoro, 2012; Tsaparina-Guillemard et al., 2011).

Second, we also investigated the effect of context typicality, defined as the global scene the object appears in. Contrary to our expectation, the results suggest that context typicality does not have an effect on naming variation in naming tasks of descriptive nature. We speculate about two possible explanations for this result; further research is needed to elucidate the role of context typicality. On the one hand, it may be that, contrary to our findings, there is an effect but we fail to detect it. This may be 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); and the effectiveness of Anderson et al. (2018)'s model in extracting relevant visual features from images is additionally confirmed by the fact that our results for object typicality corroborate previous findings. That being said, whether our context representations are informative enough to obtain good estimates of context

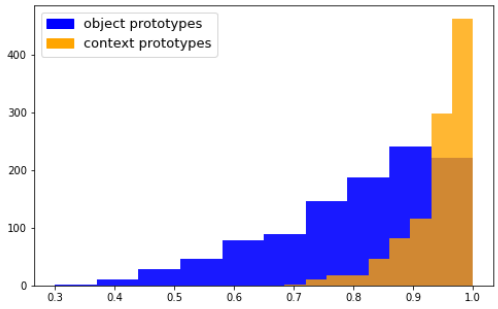


Figure 6: Histograms of prototypes' similarity between top name - alternative name pairs.

typicality remains an open question.

A second possibility may be the case that context – construed as global scene– truthfully does not affect naming variation in a descriptive task such as the one the ManyNames subjects were asked to do. Here, the nature of the task may be crucial: When asked to freely produce a name for an object, speakers may not be influenced by the visual properties of the scene. This may, at least in part, be due to prototypical contexts for top names and alternative names often being very similar. For instance, the names "armchair" and "chair" are often naming alternatives for the same object, but the prototypical context for the two names may be similar: both armchairs and chairs typically appear in living rooms. We looked into this possibility by computing the cosine similarity between the prototypes of top and alternative names. We found that the similarity between the prototypes of top names and alternative names indeed tends to be higher for context prototypes (M=0.94, SD=0.05) than for object prototypes (M=0.81, SD=0.14). Figure 6 shows the entire distributions. If the contexts for naming alternatives are similar, they likely do not provide much information to make naming preferences go one way or the other.

Note that, in this scenario, it is still possible that the typicality of the visual context plays a role for naming phenomena other than variation, such as naming speed. For instance, it could be that producing a name for an object that appears in a very atypical context demands more time, devoted to recognizing the object and choosing a name for it.

Lastly, we have specifically defined context as the *global scene* in which an object is embedded. The analysis of other aspects of context may yield different results; for instance, distractor objects in the scene competing with the target may affect naming even in a descriptive task (recall that distractor objects play a role in naming choice in referential tasks; (Graf et al., 2016; Jescheniak et al., 2005)). Past research suggests that this may be the case, at least to a certain extent (Van Der Wege, 2009).

Beyond further probing the role of context typicality in naming, analyzing the properties of *all* alternative names, as opposed to only the most frequent alternative, as we have done here, is a promising venue for future research. The opposite effects on variation of top name and alternative name properties seem to suggest that a competition between names in the lexicon takes place when speakers have to name an object (for a review see Spalek et al., 2013). In further work, we plan to run a new analysis on the ManyNames dataset, designed for the purpose of taking into account all the objects and all the names associated to them, shedding light on this competition: When multiple name alternatives are similarly good name candidates for an object, we expect variation to increase; when only one name fits the object, we expect speakers to agree more easily on the object name.

In sum, we have presented a large-scale computational analysis using naturalistic stimuli to investigate sources of naming variation. Our results confirm previous findings when it comes to the most frequently used name for an object: the more visually typical the object is of it, and the more frequent the name, the less naming variation it evokes. They also offer novel insights when it comes to the role of objects' alternative names and context typicality, suggesting a competition between names but not context effects. Finally, we demonstrate that state-of-the-art computational models can provide helpful methods to address open research questions, and to corroborate previous findings on a larger scale.

Acknowledgements

The authors thank the reviewers and 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 the Spanish Research Agency (ref. PID2020-112602GB-I00). 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

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.

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*.

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, 09). Bank of Standardized Stimuli (BOSS) phase II: 930 new normative photos. *PloS one*, *9*, e106953.

Brysbaert, M., & New, B. (2009, 11). Moving beyond kucera and francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english. *Behavior Research Methods*, *41*, 977-90. doi: 10.3758/BRM.41.4.977

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

Glaser, W. R. (1992, January). Picture naming. *Cognition*, *42*(1-3), 61–105. doi: 10.1016/0010-0277(92)90040-O

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*.

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

Humphreys, G. W., Riddoch, J., & Quinlan, P. T. (1988). Cascade processes in picture identification. *Cognitive Neuropsychology*, *5*(1), 67-104.

Jescheniak, J., Hantsch, A., & Schriefers, H. (2005, 10). 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*(1), 13–139.

Jolicoeur, P., Gluck, M. A., & Kosslyn, S. M. (1984). Pictures and names: Making the connection. *Cognitive Psychology*, *16*(2), 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., ... Li, F.-F. (2017, 05). Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, *123*.

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

Levelt, W. J. M., Schriefers, H., Vorberg, D., Meyer, A. S., & et al. (1991). The time course of lexical access in speech production: A study of picture naming. *Psychological Review*, *98*(1), 122–142. doi: 10.1037/0033-295X.98.1.122

Liu, Y., Hao, M., li, P., & Shu, H. (2011, 01). Timed picture naming norms for mandarin chinese. *PloS one*, *6*, e16505. doi: 10.1371/journal.pone.0016505

Moreno-Martínez, F., & Montoro, P. (2012, 05). An Ecological Alternative to Snodgrass & Vanderwart: 360 High Quality Colour Images with Norms for Seven Psycholinguistic Variables. *PloS one*, *7*, e37527.

Ren, S., He, K., Girshick, R., & Sun, J. (2015, 06). Faster r-cnn: Towards real-time object detection with region proposal networks. In *Ieee transactions on pattern analysis and machine intelligence* (Vol. 39).

Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., & Savarese, S. (2019, June). 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., 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.

Silberer, C., Zarrieß, S., & Boleda, G. (2020). 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.

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.

Spalek, K., Damian, M. F., & Bölte, J. (2013, June). Is lexical selection in spoken word production competitive? Introduction to the special issue on lexical competition in language production. *Language and Cognitive Processes*, *28*(5), 597–614. doi: 10.1080/01690965.2012.718088

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-Guillemard, D., Bonin, P., & Méot, A. (2011, 06). 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

Van Der Wege, M. M. (2009). Lexical entrainment and lexical differentiation in reference phrase choice. *Journal of Memory and Language*, *60*, 448-463.

Vitkovitch, M., & Tyrrell, L. (1995). Sources of disagreement in object naming. *The Quarterly Journal of Experimental Psychology Section A*, *48*(4), 822-848. doi: 10.1080/14640749508401419