Context matters: evaluation of target and context features on variation of object naming

Nikolai Ilinykh and **Simon Dobnik**
Centre for Linguistic Theory and Studies in Probability (CLASP),
Department of Philosophy, Linguistics and Theory of Science (FLoV),
University of Gothenburg, Sweden
name.surname@gu.se

Abstract

Semantic underspecification in language poses significant difficulties for models in the field of referring expression generation. This challenge becomes particularly pronounced in setups, where models need to learn from multiple modalities and their combinations. Given that different contexts require different levels of language adaptability, models face difficulties in capturing the varying degrees of specificity. To address this issue, we focus on the task of object naming and evaluate various context representations to identify the ones that enable a computational model to effectively capture human variation in object naming. Once we identify the set of useful features, we combine them in search of the optimal combination that leads to a higher correlation with humans and brings us closer to developing a standard referring expression generation model that is aware of variation in naming. The results of our study demonstrate that achieving human-like naming variation requires the model to possess extensive knowledge about the target object from multiple modalities, as well as scene-level context representations. We believe that our findings contribute to the development of more sophisticated models of referring expression generation that aim to replicate *human-like* behaviour and performance. Our code is available at https://github.com/GU-CLASP/object-naming-in-context.

1 Introduction

The adaptability of human language presents a significant challenge for computational modelling, as it relies on both external contextual factors and internal personal beliefs and goals of the language users. The significance of the intents and goals cannot be overstated, as they dictate the specific choice of referring expressions and object descriptions (van Miltenburg, 2017; Ilinykh et al., 2018; Alikhani and Stone, 2019; Baltaretu et al., 2019; Mädebach et al., 2022). Furthermore, these choices can vary depending on the specific task or the absence thereof. Put simply, language continues to evolve and adapt, while existing models are typically trained to generalise. Evaluating such systems proves hard, as evaluation metrics typically assume a single optimal solution, disregarding other valid alternatives (Kreiss et al., 2022). As variation in language arises due to different levels of underspecification between language units (words) (Pezzelle, 2023), addressing this problem brings valuable insights into understanding the effects of the task, contexts and how their interplay can be modelled.

But what is the "task"? And how do we define "context"? A task-oriented language use is often understood through the prism of human-human interaction, where communicative goals are important (Brennan and Clark, 1996). During these interactions, a shared understanding, known as a common ground, is established to optimise communication (Stalnaker, 1978). What ends up being in common ground is dependent on the task, and the importance of tasks and intents for modelling language has been emphasised in many recent proposals to language grounding (Andreas, 2022; Schlangen, 2022; Giulianelli, 2022; Fried et al., 2023). In contrast, language can be used to simply describe objects in the world with an intent to **identify** them. These intents are typically determined by the set of instructions provided to a human e.g. "describe an image" (Lin et al., 2014). In doing so, we perform *the object identification task* which is a communicative act, albeit a highly specific one.

The intent to simply describe things without a specific communicative goal has been one of the traditional tasks in the field of natural language generation (NLG). As referring is an important aspect of human communication (Frank and Goodman, 2012), much computational work has focused on building automatic referring expression generation systems (Krahmer and van Deemter, 2012). The primary goal of referring expression generation is

Proceedings of the 1st Workshop on Linguistic Insights from and for Multimodal Language Processing, pages 12–24

to produce a text in natural language that identifies a target object within a given context (Reiter and Dale, 2000) by making the object uniquely identifiable from the distractors. In the absence of the communicative intent, the definition of "given context" becomes extremely important as it directly influences referring (Schüz et al., 2023). **Visual context**, for instance, plays a crucial role in determining the content of the referring expression. This can be exemplified by multiple variables such as naturalness of the scenes where the target object appears (van Deemter et al., 2006; Mitchell et al., 2013; Kazemzadeh et al., 2014) or the presence of visual distractors and their position relative to the target object (Graf et al., 2016) and the typicality of the visual context as a whole (Gualdoni et al., 2022a,b,c). But visual context is not the only context available in the task of referring. Humans also rely on their knowledge of the world when describing things, and their **background knowledge** influences the choice of referring given a specific visual context (Dale and Viethen, 2009). In fact, the use of various names to refer to a single entity stems from the fact that different speakers tackle underspecification in different ways. Humans use given context to fill in the missing information, but they do so differently based on individual perspectives. Therefore, investigating the effect of different contexts on the naming variation and capturing human behaviour in models is beneficial for developing a better REG architecture.

This study addresses two challenges: (i) existing models of referring are simply not learning to approximate possible names for entities and (ii) it is hard to generate a correct name if the level of semantic underspecification is high. As underspecification is correlated in humans with variation, we assume that the models that approximate human behaviour should be equally "confused" as humans when generating descriptions and should produce the same variation. For a model that is behaving this way we can be sure that the variation is due to the way they capture semantic knowledge and context sensitivity rather than the noise (e.g., better performance on more frequent labels). **Our primary questions** are as follows: what is the set of features that enables computational model to closely capture the variation observed in human object naming? Can we combine such features to get closer to a REG model that can capture human-like object naming?

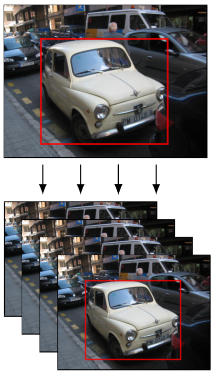
To address the questions outlined above, we investigate the effects that different context representations have on the model that is tasked with predicting an object name. We use CLIP (Radford et al., 2021) to encode different context representations and train a simple classifier to predict target object names using the Many Names dataset (Silberer et al., 2020b,a). We specifically examine how different features influence model's ability to capture human object naming variation. Through the comparison of the model's performance with humans across various metrics, we identify features that assist the model in making more valid and contextually motivated approximations of naming variation, reminiscent of human behaviour. We then combine different features and examine their fit for capturing naming variation. Our results demonstrate that the model that captures contextual sensitivity of object naming well (be it language or vision or both) is a good approximation of human knowledge and behaviour. We note that, unlike Silberer et al. (2020b), we are testing how different types of knowledge contribute to naming variation rather than building or evaluating object naming models. While Silberer et al. (2020b) also focus on typicality and whether the name is the top one or an alternative one in naming, we are interested in individual variation and the effects of context representations on the "distortions" of such typicality.

2 Problem formulation

2.1 Dataset

As our dataset, we use the Many Names dataset (Silberer et al., 2020b) as it provides a suitable testbed for studying naming variation. This dataset stands out from other language-and-vision data collections that can be used for studying naming variation (Mitchell et al., 2013; Kazemzadeh et al., 2014; Plummer et al., 2015; Yu et al., 2016; Krishna et al., 2017) due to its high number of name types per object and alignment between names and objects. This way we can directly study the variation in reference to entities. The dataset was created by picking a single target object per image based on annotated data from Visual Genome (Krishna et al., 2017). Next, name annotations for each object were collected from multiple crowd-workers[1]. There are on average 36 name tokens per object in Many Names, and their name types are sorted based on the frequency of being used to refer to

[1]For details, see Silberer et al. (2020a).



{car: 32,
vehicle: 2,
automobile: 1}

{car, car, car, ...,
vehicle, vehicle,
automobile}

Figure 1: Dissecting the Many Names dataset (Silberer et al., 2020b) into individual instances. The `Target` condition is depicted in which the model was provided with features of the object in the red box; datasets for `Context-Obj` and `Context-Scene` were built in the same way.

objects. An example from the Many Names dataset is shown in the upper part of the Figure 1. In our experiments, we use the dataset splits of Many-Names v2.1 as reported in Silberer et al. (2020b). Specifically, the train / val / test splits consists of 21503 / 1110 / 1072 items respectively.

2.2 Learning scheme

We approach object naming through the prism of referring expression generation. Our objective is to capture human-like variations in naming. Therefore, we shall look into the probability distribution of names that the model produces in a given context. Training a model to approximate naming distribution similar to humans should improve referring expression generation, possibly reducing deterministic nature of the models (van Deemter et al., 2012). However, one problem with the naming distribution in model's output is that it may include invalid or non-human-like naming variations. To address this, we aim for our models to demonstrate shifts in the probability distribution, mirroring the changes observed in human object naming. These shifts are then learned by mapping different representations corresponding to visual context and background knowledge, rather than random noise, with the target names.

While it is possible to build different models per speaker to account for variation among these speakers (Dale and Viethen, 2009), our goal is to develop *a single function* that can approximate such variation across multiple individual describers. We deliberately chose to train such a simple model because it allows us to focus on evaluating the contribution of features to naming variation rather than the model's complexity. We ask if this function can predict the likelihood of a speaker referring to a particular object with a particular name. To answer the question, we break down the individual accumulated counts of frequencies into the number of individual referring events, each consisting of one description. This approach is similar to that of Coventry et al. (2005). The frequency of these events in the dataset reflects the likelihood that the object would be referred to with that name. The bottom part of Figure 1 provides a more detailed example, which involves breaking down the counts of different name types from individual instances. This mirrors how humans describe an image, where each person may use different names for the same object. By learning from these individual instances, the network is expected to learn the variations in naming and, therefore, capture speaker uncertainty. During training, the model is repeatedly presented with input–"car" pair 32 times, while inputs mapped with "vehicle" and "automobile" are shown to the model 2 and 1 time, respectively. This variability in selection is akin to the diverse choices humans make in object naming. By using such training scheme, we encourage the model to learn *uncertainty* inherent in human naming, which is important for capturing variation. In the next section, we will describe how we represent different inputs to the name prediction model.

2.3 Input representation

The dataset consists of the following elements: for the j^{th} sample, there is an image i_j, a target object t_j with a bounding box t_j^{bb} obtained from Visual Genome, and a dictionary V_j containing names and their frequencies assigned to t_j by crowd-workers. Our initial proposal is to use each feature independently as input to a simple classifier to evaluate individual contribution of features. Next, a combination of different features can be explored. In terms of the features, we examine different types of representations which differ in the level of contextual information available. These include features that solely focus on the target object (`Target`), features that incorporate information about surrounding objects but exclude the target object (`Context-Obj`), and features that cover knowledge about the entire scene (`Context-Scene`). For each feature type, we consider three representation modes: visual, lin-

guistic, and their combination. We encode each feature type with CLIP (Radford et al., 2021)[2], a pre-trained multi-modal transformer that learns strong multi-modal representations through its contrastive learning on large amount of image-text pairs. Our motivation for selecting different modalities and combining them is as follows. Text features can be seen as representations of the background knowledge in terms of the meaning of a word in the contexts that were given to the pre-trained model, e.g. CLIP. This knowledge is acquired through extensive pre-training, and CLIP, in particular, possesses rich contextual information about entities and objects. Hence, textual features encode *general* knowledge about the interaction of these objects, not related to particular events (although it is possible that due to naming variation of labels some specific local context is also captured). An example of this type of world knowledge includes the typical contexts in which bananas appear (kitchen, food, nature, market), how they are typically used (eaten, consumed), and who typically uses them (humans, animals). On the other hand, vision features contain information about the immediate context of the target object. Their purpose is to encode the situation in which the object appears in a specific case. Here is an example of this type of feature: a more detailed and specific understanding of the situations in which bananas appear could involve a market with various fruits of different colours and a better understanding of how bananas fit into this specific context. By integrating both these feature types, we take a step toward modelling the information sources that humans employ for object naming. These features include world knowledge about how objects interact in the world and specific visual information about these objects.

In the `Target` condition, our aim is to examine the effect of the knowledge about the target object in the process of object naming. We seek to determine whether a model can effectively capture naming variation in the absence of contextual information, relying solely on the appearance and/or common sense knowledge of the target object. To represent common sense knowledge[3], we use labels that have been assigned to objects (both target and context) by the annotators of the Visual Genome dataset (Krishna et al., 2017). By encoding these labels with CLIP, we can leverage strong signals and extensive additional knowledge about the objects. It is important to note that this type of information is not typically available to a conventional referring expression model. In fact, any identification system that uses this information would be considered cheating in predicting names. In our experiments, we incorporate this knowledge to evaluate its contribution to generating a variety of names, but it is important to acknowledge that this feature may or may not be available in individual tasks.

With the `Context-Obj` condition, we measure how well a target's name can be predicted from surrounding objects alone. In other words, can we "guess" a name based on the visual and/or common sense knowledge about context objects? Finally, with the `Context-Scene` condition, we focus on attention and search: given visual and/or common sense knowledge about the scene as a whole (e.g., all objects treated equally, no difference between context or target objects), can we model human naming variation?

`Target` We represent visual $\mathbf{v}_j^v$ and linguistic $\mathbf{v}_j^\ell$ information about the target object as follows:

$$\mathbf{v}_j^v = f_{\text{CLIP}}\,(t^{\text{bb}}j), \tag{1}$$

$$\mathbf{v}_j^\ell = f_{\text{CLIP}}\,(t_j^{\text{VisGen}}). \tag{2}$$

Here, t_j^{VisGen} represents the label of the target object from Visual Genome.

`Context-Obj` Another type of feature that can be explored is the knowledge of context. In this particular setup, the input representations do not contain any information about the target object, whether visual or common sense-related. This setup can be viewed as a "guessing game" where the model is given a context representation and tasked with predicting the name of an object likely to appear in that context. To model this scenario, we use Visual Genome annotations to represent the context of the target object. Specifically, we extract a list of bounding boxes for all objects that are *not* the target object, denoted as $\mathbf{R}_{\setminus t_j} := (r_1, \ldots, r_K)$, where K is the number of objects in i_j. Then,

$$\bar{\mathbf{v}}_j^v = f_{\text{CLIP}}\,(\mathbf{R}_{\setminus t_j}), \tag{3}$$

$$\bar{\mathbf{v}}_j^\ell = f_{\text{CLIP}}\,(\mathbf{L}_{\setminus t_j}), \tag{4}$$

where $\mathbf{L}_{\setminus t_j}$ is the list of object descriptions, where each element is a simple phrase consisting of a

[2]We use a pre-trained ViT-L/14@336px based on the code from the official CLIP GitHub repository: https://github.com/openai/CLIP.

[3]In this study, we use the terms "linguistic" and "common sense" interchangeably, as they both refer to the knowledge and understanding of language-related information and general knowledge about the world.

name and up to five attributes from Visual Genome annotations, e.g. "car black big", and $\bar{\mathbf{v}}$ is the average of the objects or their descriptions. We also apply L2 normalisation on the resulting vector to obtain a more robust context representation. This normalisation helps enhance the discriminative power of all feature vectors and disregards the influence of differences in magnitude and scale[4]. The motivation behind this design choice is further described in Appendix A.

`Context-Scene` In the third experiment, our focus is to examine the predictability of naming variation from the context *as a whole*. We use perceptual features of the entire image that have been encoded with CLIP and incorporate object-relation triplets that describe the content of the scene. These triplets are sourced from the Visual Genome dataset, where each image is annotated with relationships. We note that that these relationships are generated by different crowd-workers, ensuring a diverse range of annotations for our experiment. While the number of relations may differ from image to image, they collectively provide an overview of the objects present in the scene and their associated events. By leveraging these relationships, we can create language input features for the `Context-Scene` model:

$$\mathbf{v_t} = f_{\text{CLIP}}\left(<\text{S},\text{P},\text{O}>\right), \tag{5}$$

where $<\text{S},\text{P},\text{O}>$ represents a single string comprising the subject, predicate, and object names of a specific relationship triplet. Since annotated scene contexts in Visual Genome are not predetermined and vary across images, textual descriptions can be constructed in various ways. To generate textual scene descriptions, we shuffle and randomly extract a varied number of relationship strings. We then employ different methods to feed these strings to the CLIP model in order to obtain language features. Subsequently, we evaluate the `Context-Scene` model using each type of text representation to identify the one that demonstrates optimal performance. The selected model is then used in our primary experiments. More details on how the best `Context-Scene` model that uses text was chosen can be found in Appendix B.

[4]In each experiment where we need to create a single vector from a list of vectors, our approach is to first compute the average vector from the list and then normalise it.

3 Model

In this study, we adopt a simple approach by constructing a CLS (classification) model. The objective is to approximate a function that can predict naming variation. The success of this function approximation provides insights into the suitability of the features as predictors of naming variation. The approach is akin to the use of generalised linear models in statistical testing, where we aim to capture the relationships between the features and the predicted labels. To maintain a close connection to linearity, we build a single-layer feed-forward network as our model. We specifically examine the probabilities assigned to all the labels predicted by the model and evaluating their degree of variation against the probabilities assigned by humans.

The model is trained following the scheme outlined in Section 2.2 and takes input representations described in Section 2.3. The model takes $\mathbf{x}$ which is either a concatenation of visual and linguistic features $\mathbf{x} = (\mathbf{v}_v \oplus \mathbf{v}_\ell)$ or a uni-modal feature, e.g. $\mathbf{x} = \mathbf{v}_v$ or $\mathbf{x} = \mathbf{v}_\ell$, where $\mathbf{x} \in \mathbb{R}^{1\times 768}$. The model is trained to predict a target name $\mathbf{y}$ from the set of all possible names that are available: $\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$, where $N = 1642$ is the number of all possible names. N is determined by the set of unique names across all data splits. The model is defined as follows:

$$\hat{\mathbf{y}} = \sigma\big(\left(f_2\left(f1\left(\mathbf{x}\right)\right)\right)\big), \tag{6}$$

where

$$f_1(\mathbf{x}) = \text{ReLU}\left(\text{BN}\left(\mathbf{W}_1\,\mathbf{x} + \mathbf{b_1}\right)\right), \tag{7}$$
$$f_2(\mathbf{x}') = \text{Dropout}\left(\mathbf{W}_w\,\mathbf{x}' + \mathbf{b_2}\right) \tag{8}$$

where $\mathbf{W}_1 \in \mathbb{R}^{d_1\times d_2}$, and $\mathbf{W}_2 \in \mathbb{R}^{d_2\times 1}$ is output linear layer that produces the list of logits $\tilde{Z} \in \mathbb{R}^{1\times N}$. The model applies $\mathbf{softmax}\ \sigma$ over the last dimension of $\tilde{Z}$ to transform unnormalised scores into name probabilities.We adjust d_1 depending on the type of the experiment: if we test features from a single modality, then $d_1 = 768$, otherwise $d_1 = 1536$. We set $d_2 = 512$ and $\text{Dropout} = 0.1$.

All models were trained using a batch size of 64 and standard cross-entropy loss. The Adam optimiser (Kingma and Ba, 2015) with a weight decay of $1e-5$ was used, and the learning rate was set to $4e-3$. During training, the gradients were clipped by their norm per single batch, with a maximum norm set to 3. The models were trained

for a total of 200 epochs, and the best model was selected based on the validation loss at the epoch level. Additionally, we used a scheduler, reducing the learning rate if there was no improvement in the loss for three consecutive epochs during validation.

4 Evaluation metrics

To evaluate the general performance of the model, we use multiple metrics. We note that during evaluation, we do not differentiate between top and alternative names. Our model learns that each possible name is valid but to varying degrees based on the frequency of being assigned to an object. The model is never presented with multiple names and their frequencies simultaneously. This means that it does not make comparative judgments about one name being more or less valid than another. Therefore, our results should be interpreted as an assessment of how often the model would use a specific name to describe an object, without considering its relation to other alternatives.

Firstly, we measure the model's ability to predict the top name (e.g., the most frequent name) by looking at accuracy @1. Other degrees of accuracy are also useful to consider, as they indicate whether the top name occurs in the top-*k* predictions generated by the model, where *k* is the number of name types used to describe a specific target in the specific image. The final accuracy scores are reported as averages over the total number of samples. We also compute the mean rank of the ground-truth label among the model's predictions and report the average mean rank (AMR) across all items. Additionally, we measure the perplexity of the models as an indicator of overall predictive performance. Unlike accuracy, which solely focuses on comparing the top name, perplexity allows us to compare the variation in the predictions of different names. However, perplexity does not measure semantic equivalence or similarity between the predicted names and the human-generated names. We note that since we have previously evaluated the success of the model with accuracy, we can assume that such noise is minimised. We compute perplexity $\mathbf{PP}$ by taking the logarithmic base of the entropy and raising it to the power of entropy, e.g. $\mathbf{PP} = \exp^{\mathbf{H}}$.

To evaluate the suitability of features for predicting naming variation, we calculate the entropy (Shannon, 1948) of each model and humans. Entropy helps us quantify uncertainty, and we anticipate that the best model will demonstrate a similar level of uncertainty as humans. To assess the degree of association between the entropy of each model and human responses, we compute Spearman's rank correlation coefficient (Spearman, 1904). This metric measures the monotonic relationship between the two, and it serves as our primary evaluation metric. The way entropy is calculated is slightly different between the model and humans in terms of the probabilities that we use. For the model, we take the degree of belief that the object should be assigned a particular label by the neural network, represented by logits $\tilde{Z}$. These logits are transformed into probabilities using the $\mathbf{softmax}$ function: $\mathbf{P}_m = \sigma(\tilde{Z})$. For humans, we consider the probability (derived from frequencies) that a human would assign a particular label to the object, representing a collective likelihood. For each test item, we collect all available ground-truth human responses (m) and their corresponding frequencies ($x_1, x_2, \ldots, x_m$). These frequencies are then transformed into probabilities:

$$p_i = \frac{x_i}{\sum_{j=1}^{m} x_j}, \quad \text{for } i = 1, 2, \ldots, m. \tag{9}$$

Next, we construct a new vector $\mathbf{P}_h \in \mathbb{R}^{1\times N}$, where values in positions corresponding to the positions of each response in the model's dictionary $\mathcal{V}$ (with $|\mathcal{V}| = N$) are replaced with their respective probabilities p_i, and the rest are set to 0. To compute entropy $\mathbf{H}$ of $\mathbf{P}m$ and $\mathbf{P}h$, we use the following operation:

$$\mathbf{H}_{m\setminus h} = -\sum_{k=1}^{|\mathbf{P}_{m\setminus h}|} p_k \log p_k. \tag{10}$$

We normalise the maximum attainable entropy by $-\log\exp(N)$ to ensure comparability between different models, resulting in entropy values ranging between 0 and 1, where 1 represents the highest possible entropy. All metrics are reported as averages across the test set. We anticipate that the model probabilities will show greater variation across labels due to noise compared to humans, as the model may assign low probabilities to labels that are not applicable. On the other hand, humans tend to produce "cleaner" labels as they are direct judgments. To address this issue, we compare the ranks of entropies using correlation coefficients. This choice is relevant because the

	Condition	Mode	Accuracy (%) ↑			AMR ↓	PP ↓	H ↓	ρ
			@1	@5	@10				
1	Target	TEXT	69.15	87.68	89.94	41.45	4.745	0.210	0.540*
2		VISION	56.70	81.09	86.34	52.87	7.199	0.266	0.485*
3		VISION-TEXT	70.02	90.99	92.30	33.77	3.740	0.178	0.574*
4	Context-Obj	TEXT	40.90	67.58	76.73	52.13	14.924	0.365	0.343*
5		VISION	49.14	75.14	83.20	40.79	10.360	0.315	0.328*
6		VISION-TEXT	46.48	72.98	81.04	45.87	11.531	0.330	0.321*
7	Context-Scene	TEXT	4.09	16.85	31.80	59.00	51.111	0.531	-0.024
8		VISION	47.93	73.51	81.42	60.73	9.116	0.298	0.410*
9		VISION-TEXT	53.34	77.91	83.98	38.87	8.281	0.285	0.424*
	Human						**1.623**	**0.065**	**1.000**

Table 1: Evaluation of different features (models 1-9) against human scores. We highlight the top three models **per condition** in each metric, with colour intensity reflecting their performance (stronger indicates better). Human scores are provided as a reference. The values of Spearman correlation ρ with *denote a very high level of significance, e.g. p-value ≤ 0.001.

vector $\mathbf{P}_h$ contains many zero values, which motivates us to focus on the ranks of the values rather than the values themselves. When describing an object, humans select from a limited set of "valid" names, whereas the model considers both "valid" and "invalid" names (a total of 1642 possible name types). By examining the ranks of the model's predictions, we mitigate this issue. We would like to emphasise the general importance of statistical testing to determine the extent to which the model's performance is influenced by either the network design or the features themselves. In this paper, we employ Spearman correlation to measure the relationship between input features and target variables. This test is appropriate because we are interested in whether the simple neural network can approximate a function between input features and the resulting naming variation. This correlation shows whether there is a linear relation between the model's prediction and human scores and, therefore, whether those input features are associated with human scores. We believe that future work can focus on measuring the effects not only of features but also of the model's design on naming variation.

5 Results

Table 1 demonstrates the results of our experiments, which focused on evaluating different feature representations (modes) for various feature types (conditions) in modelling naming variation. Firstly, we examine differences within each condition and analyse different modes to identify the best features for representing specific condition. Next, we explore the differences between conditions and consider the potential of combining them to achieve a more human-like performance in the object identification model. We conclude by emphasising features that need to be encoded by an REG (Referring Expression Generation) model to effectively capture human-like object naming variation.

5.1 Best feature per condition

Representing targets In the Target condition, multi-modality proves to be crucial as it achieves the highest performance in predicting the correct answer, exhibiting the lowest mean rank and perplexity. Additionally, language-and-vision features significantly reduce uncertainty and bring it closer to human levels, as indicated by entropy and correlation measures. Notably, language appears to contribute more to the fusion of modalities, as it offers greater informativeness compared to visual information. This observation aligns with previous studies conducted on various multi-modal tasks (Agrawal et al., 2018). The contribution of the text mode can be attributed to the degree of semantic similarity that an object label from Visual Genome and a target name share with each other. For example, the Visual Genome label for the target object in Figure 1 is "sedan", which is very similar in meaning to the target names, while context labels ("street", "human") might be less useful

in reducing uncertainty for naming. Additionally, encoding it with CLIP that is expected to understand relations between "car", "sedan" and "vehicle" might provide even more informative representations, reducing ambiguity about the choice of the name. Nonetheless, the vision representation in the `Target` condition demonstrates good performance, as it does not lag far behind the performance of the text features. One possible explanation for this result is that the knowledge in text is simply not very effective, either due to noise or its challenging nature to learn from, or it may not be very informative. We emphasise that it is important to evaluate the quality of knowledge types in the Limitations section. Interestingly, incorporating visual appearance of the target object further enhances the correlation between the predicted and human naming variation. We conclude that for effectively representing the target object, the most optimal feature representation involves combining visual information with common sense knowledge of the target object.

Representing context as objects In the `Context-Obj` condition, the vision-only model demonstrates the best performance in predicting a single correct name and achieves the lowest mean rank of the correct name in its predictions. It also has the lowest entropy among the different modes considered. However, it is important to note that the vision-only model does not exhibit the highest correlation with human naming variation. The highest correlation is observed when the model relies solely on textual features, despite having the highest entropy among all three modes. This observation is interesting as it emphasises the significance of world knowledge in capturing naming variation. Understanding what objects might co-occur in a given context provides valuable information to the model (Dobnik et al., 2022). For instance, having the context labels "counter", "fridge", and "oven" might assist the model in predicting the target name "pot" more accurately than relying solely on visual features of these context objects. Interestingly, contrary to the `Target` condition, combining linguistic and visual information leads to the lowest correlation score. Based on these results, we conclude that representing context in a model that aims to capture naming variation is best achieved through the textual labels of the context objects.

Representing context as a scene When representing context as a single image with or without relationship triplets, combining language and vision yields the best performance across various metrics, including correlation with humans. There is a notable reduction in uncertainty and an increase in correlation when the model has access to the visual appearance of the context alone, represented by the image as a whole. This improvement can be attributed to the model's ability to better contextualise the target object as text knowledge provides only general information about what context objects are and lacks details on how the objects actually look. In contrast, uncertainty in the model is significantly high when the model is provided with relationship triplets alone. In fact, this condition shows no correlation with human naming at all. The text-only model stands out with exceptionally high perplexity and significantly higher entropy compared to any other model in any of the conditions. We believe this highlights the importance of choosing appropriate representations for conveying textual knowledge about the scene. Exploring the performance of models using other types of representations, such as scene categories, captions, or more coherent scene descriptions, is left as a topic for future investigation. Considering that the task involves mixed representations of targets and context without explicit labelling, the `Context-Scene` model approximates correlation most effectively when there is a fusion of modalities.

Overall, the findings indicate the importance of the text modality in learning about the target object. However, combining text with vision is necessary to achieve lower entropies and higher correlations with human naming. This demonstrates that predicting a name solely from text is challenging because the model lacks knowledge about the appearance of objects and struggles to determine what to focus on. Access to visual representations allows the model to differentiate between targets and contexts, possibly due to factors such as the perspective and location of the objects, which are relevant for naming. In the next section, we focus on identifying the optimal feature combination for better object naming. Our goal is to assess the correlation with human naming when multiple conditions are combined, thereby determining the best possible combination of features.

Condition	Accuracy (%) ↑			AMR ↓	PP ↓	H ↓	ρ
	@1	@5	@10				
3+9	71.02	88.59	90.62	**37.66**	**3.773**	**0.179**	**0.580***
3+4	70.55	88.76	90.62	43.02	4.187	0.193	0.568*
3+9+4	**71.41**	**89.73**	**91.42**	38.96	3.995	0.187	0.578*

Table 2: Evaluation of different combinations of the best-performing features from Table 1. The meaning of colour intensity and *is described in Table 1. The numbers in condition correspond to the features from Table 1.

5.2 Combining best-performing features

Here we test different feature combinations to replicate human-like naming variation. We acknowledge that without testing of *all* possible combinations, we cannot really conclude which feature combination is the best. However, here we have chosen feature combinations based on our intuition regarding what is commonly found in models and what yields the best performance when considering individual features. Table 2 presents the results of combining features that have shown the highest correlation with humans across different conditions. For each condition, we progressively combined features that showed the highest correlation with humans by concatenating them together. As a result, the input vector size for the 3+9+4 condition became 5×768, representing the combination of two modalities for the target, two modalities for the context as a scene, and one modality for the context as objects. The best model, which incorporates visual and common sense knowledge about the target (3 in Table 1) along with multi-modal knowledge about the scene (9 in Table 1), achieves the lowest entropy and improves the correlation with humans compared to the previously best model, the Target model. This indicates that combining the appearance of an object, including its label, with the shared context and thematic representation of the scene as a whole can be beneficial. Interestingly, combining different features with each other generally yields better results than using them individually, except for the combination of the best Target and Context-Obj models. The optimal combination is found to be the integration of knowledge about the target with knowledge about the scene as a whole. Notably, the 3+9 combination achieves lower accuracies, suggesting that it may be more focused on capturing variation rather than predicting the most probable name. These findings have implications for the representation of context. While the visual appearance of objects is important, it also needs to be presented in a consistent and comprehensive manner, such as using a whole image where the relationships among context objects are clear, and the fit of the target within the overall context can be easily extracted.

6 Conclusions

Naming and language in general is semantically underspecified (Frisson, 2009; Pezzelle, 2023). To fill in the missing gaps in reconstructing meaning, language users rely on contextual information, be it perceptual information or background knowledge. In this study we examined different types of context representations for capturing human object naming variation. We have found that to capture naming variation it is important to have a lot of knowledge about the target object. We also have shown that the way context is represented matters: object-level visual representations might narrow down the gap in uncertainty between models and humans, but they might not correlate the most with humans in object naming. Future work on this topic should focus on using encoders other than CLIP, building more complex classifiers and investigating the effect of different ways to represent common sense knowledge (e.g., not relationship triplets, but captions or another type of image descriptions). Also, looking at object naming in a task context with communicative goal is another important direction.

Limitations

Information fusion This work uses averaging to generate a single vector when combining multiple language and/or vision features. It should be acknowledged that adopting an alternative fusion method, such as multiplication or summation, could potentially affect the final scores of the models, particularly when the differences between them are relatively minor. We recognise that the results reported in this study are specific to the particular technical setup employed, involving L2 normali-

sation with averaging. Hence, further investigation is warranted to determine whether the reported findings remain consistent when using a different fusion method. Some of our ideas for information fusion are presented in Appendix A. In addition, fusing different features from different conditions by multiplying them or learning a function to fuse them can be an alternative to a simple concatenation that we use in this study.

Knowledge representations We note that in the context of a standard REG task, knowing the label of the target is practically impossible. Hence, it is expected that a model with linguistic knowledge about the target would perform well. Also, adding more features (visual, linguistic, others) appears to hinder performance due to the increased number of parameters and a larger hypothesis space. Therefore, the objective of learning should be to strike a balance between model size and feature informativeness. It is also important to seek a knowledge representation that closely resembles how humans name objects.

Acknowledgments

The research in this paper is supported by a grant from the Swedish Research Council (VR project 2014-39) for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg.

References

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 4971–4980. Computer Vision Foundation / IEEE Computer Society.

Malihe Alikhani and Matthew Stone. 2019. "Caption" as a coherence relation: Evidence and implications. In *Proceedings of the Second Workshop on Shortcomings in Vision and Language*, pages 58–67, Minneapolis, Minnesota. Association for Computational Linguistics.

Jacob Andreas. 2022. Language models as agent models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5769–5779, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Adriana Baltaretu, Emiel Krahmer, and Alfons Maes. 2019. Producing referring expressions in identification tasks and route directions: What's the difference? *Discourse Processes*, 56(2):136–154.

S.E. Brennan and H.H. Clark. 1996. Conceptual Pacts and Lexical Choice in Conversation. *Learning, Memory*, 22(6):1482–1493.

K.R. Coventry, A. Cangelosi, R. Rajapakse, A. Bacon, S. Newstead, D. Joyce, and L.V. Richards. 2005. Spatial prepositions and vague quantifiers:: Implementing the functional geometric framework. In *Spatial Cognition IV.*, volume IV, pages 98–110, United States. Springer Nature. Error 1 : ISSN or ISBN parsed from 0302-9743 but is invalid for outputType A which is a Book.

Robert Dale and Jette Viethen. 2009. Referring expression generation through attribute-based heuristics. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009)*, pages 58–65, Athens, Greece. Association for Computational Linguistics.

Simon Dobnik, Nikolai Ilinykh, and Aram Karimi. 2022. What to refer to and when? reference and re-reference in two language-and-vision tasks. In *Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*, Dublin, Ireland. SEMDIAL.

Michael C. Frank and Noah D. Goodman. 2012. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998.

Daniel Fried, Nicholas Tomlin, Jennifer Hu, Roma Patel, and Aida Nematzadeh. 2023. Pragmatics in language grounding: Phenomena, tasks, and modeling approaches.

Steven Frisson. 2009. Semantic underspecification in language processing. *Lang. Linguistics Compass*, 3(1):111–127.

Mario Giulianelli. 2022. Towards pragmatic production strategies for natural language generation tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7978–7984, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Caroline Graf, Judith Degen, Robert X. D. Hawkins, and Noah D. Goodman. 2016. Animal, dog, or dalmatian? level of abstraction in nominal referring expressions. In *Proceedings of the 38th Annual Meeting of the Cognitive Science Society, Recogbizing and Representing Events, CogSci 2016, Philadelphia, PA, USA, August 10-13, 2016*. cognitivesciencesociety.org.

Eleonora Gualdoni, Thomas Brochhagen, Andreas Mädebach, and Gemma Boleda. 2022a. What's in a name? a large-scale computational study on how competition between names affects naming variation.

Eleonora Gualdoni, Thomas Brochhagen, Andreas Mädebach, and Gemma Boleda. 2022b. Woman or tennis player? visual typicality and lexical frequency affect variation in object naming.

Eleonora Gualdoni, Andreas Madebach, Thomas Brochhagen, and Gemma Boleda. 2022c. Horse or pony? Visual typicality and lexical frequency affect variability in object naming. In *Proceedings of the Society for Computation in Linguistics 2022*, pages 241–243, online. Association for Computational Linguistics.

Nikolai Ilinykh, Sina Zarrieß, and David Schlangen. 2018. The task matters: Comparing image captioning and task-based dialogical image description. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 397–402, Tilburg University, The Netherlands. Association for Computational Linguistics.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. ReferItGame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798, Doha, Qatar. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR (Poster)*.

Emiel Krahmer and Kees van Deemter. 2012. Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218.

Elisa Kreiss, Cynthia Bennett, Shayan Hooshmand, Eric Zelikman, Meredith Ringel Morris, and Christopher Potts. 2022. Context matters for image descriptions for accessibility: Challenges for referenceless evaluation metrics. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4685–4697, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.*, 123(1):32–73.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer.

Margaret Mitchell, Ehud Reiter, and Kees van Deemter. 2013. Typicality and object reference. *Cognitive Science*, 35:3062–3067.

Andreas Mädebach, Ekaterina Torubarova, Eleonora Gualdoni, and Gemma Boleda. 2022. Effects of task and visual context on referring expressions using natural scenes.

Sandro Pezzelle. 2023. Dealing with semantic underspecification in multimodal NLP. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12098–12112, Toronto, Canada. Association for Computational Linguistics.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2641–2649. IEEE Computer Society.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Ehud Reiter and Robert Dale. 2000. *Building Natural Language Generation Systems*. Natural Language Processing. Cambridge University Press.

David Schlangen. 2022. Norm participation grounds language. In *Proceedings of the 2022 CLASP Conference on (Dis)embodiment*, pages 62–69, Gothenburg, Sweden. Association for Computational Linguistics.

Simeon Schüz, Albert Gatt, and Sina Zarrieß. 2023. Rethinking symbolic and visual context in referring expression generation. *Frontiers in Artificial Intelligence*, 6.

C. E. Shannon. 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423.

Carina Silberer, Sina Zarrieß, and Gemma Boleda. 2020a. Object naming in language and vision: A survey and a new dataset. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5792–5801, Marseille, France. European Language Resources Association.

Carina Silberer, Sina Zarrieß, Matthijs Westera, and Gemma Boleda. 2020b. Humans meet models on object naming: A new dataset and analysis. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1893–1905, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Charles Spearman. 1904. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101.

Robert Stalnaker. 1978. Assertion. *Syntax and Semantics (New York Academic Press)*, 9:315–332.

Kees van Deemter, Albert Gatt, Roger P.G. van Gompel, and Emiel Krahmer. 2012. Toward a computational psycholinguistics of reference production. *Topics in Cognitive Science*, 4(2):166–183.

Kees van Deemter, Ielka van der Sluis, and Albert Gatt. 2006. Building a semantically transparent corpus for the generation of referring expressions. In *Proceedings of the Fourth International Natural Language Generation Conference*, pages 130–132, Sydney, Australia. Association for Computational Linguistics.

Emiel van Miltenburg. 2017. Pragmatic descriptions of perceptual stimuli. In *Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. 2016. Modeling context in referring expressions. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II*, volume 9906 of *Lecture Notes in Computer Science*, pages 69–85. Springer.

A Fusing features

In our approach, when it is necessary to combine multiple uni-modal or multi-modal representations into a single vector, we use averaging of features. This averaging process is followed by an L2 normalisation step, which normalises the features based on the Euclidean distance between individual points. Additionally, we have experimented with using multiplication for feature fusion, particularly in cases where we want to emphasise joint features or attributes and assign more importance to overlapping information. Multiplication is expected to highlight specific features that are shared across objects, such as in the case of visual features. However, we have observed that multiplication often leads to many zero values in the resulting features, and in some cases, it even leads to $\inf$ or NaN values due to the sparsity of visual representations. This sparsity can make the resulting vector difficult to learn from, especially depending on the number of objects being multiplied. Although summation of features is a straightforward approach, we have concerns that using this method results in a diluted final vector. As a result, we decided to use averaging followed by L2 normalisation as it tends to be a more effective and stable approach for feature combination.

B Representing language for `Context-Scene`

Table 3 presents the performance of various variations of the `Context-Scene` model, which incorporates the textual modality. The text representation can be either a single string containing 10 or 5 relations present in the image (`10-string` and `5-string`), or a list of different relations (`10-list` and `5-list`). The best model is selected based on the loss and average mean rank score, both computed on the test set. The best-performing model is highlighted in bold in the table.

Condition	Text Format	Accuracy (%) ↑			AMR ↓	Loss ↓
		@1	@5	@10		
Context-Scene + Text	10-list	4.04	17.98	30.43	62.63	4.774
	10-string	4.09	16.85	31.80	**59.00**	**4.676**
	5-list	3.83	16.69	31.70	63.32	4.756
	5-string	3.58	16.99	30.80	125.50	5.722
Context-Scene + Vision-Text	10-list	52.40	75.58	83.09	45.49	2.490
	10-string	53.27	77.27	83.24	43.38	2.463
	5-list	52.44	76.68	83.11	45.12	2.475
	5-string	53.34	77.91	83.98	**38.87**	**2.403**

Table 3: Performance of different Context-Scene models, which use textual modality as part of their input.