What to refer to and when? Reference and re-reference in two language-and-vision tasks

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

Abstract

How do we refer to scene entities in interactive language-and-vision tasks? We explore reference and re-reference in two tasks, link them to a model of attention and discuss our findings in relation to modelling situated interaction.

1 Introduction

In this paper we examine how conversational partners refer to scene entities in two language-and-vision tasks. Knowing the strategies and models of referring is crucial for natural language processing tasks of situated interaction, both interpretation and generation of referring expression. In natural language generation, the step is crucial for content selection (Deemter, 2016): a visual scene may include several entities, their features and spatial relations between them but only some are selected and included in the expression to be generated. In natural language understanding, referring expressions have to be resolved to scene entities, their attributes and spatial relations between them: similarly, referring expressions are ambiguous and they may be resolved to several potential candidates. In situated interaction involving several conversational partners several aspects of referring are relevant (Byron, 2003). Firstly, elements in the scenes are described to and referred to in a particular order which is reflected in the discourse model (Grosz and Sidner, 1986; Ilinykh and Dobnik, 2020; Takmaz et al., 2020). The same discourse elements may be re-referred during the discourse which is described by co-reference (Stede, 2011; Poesio et al., 2018; Loáiciga et al., 2021). When referring to discourse entities conversational participants may also take different spatial perspectives (Maillat, 2003). Our hope is that this investigation will shed light on strategies that need to be taken into consideration in modelling situated discourse. This is particularly relevant for multi-modal neural networks as understanding the properties of visual interaction will help us to evaluate and study these models for such properties (Ilinykh and Dobnik, 2022).

The mechanisms driving linguistic reference, connecting words with the physical properties of the scene, are driven by the notion of *attention*. Attention can be of two different kinds: linguistic and perceptual (visual) attention. Objects attain linguistic salience (i) if they have been mentioned in the conversation before, and (ii) depending on how thematically they are relevant to the topic of conversation and the task that the participants are engaged in. Objects attain visual salience by attention on the visual properties of the scene such as colour, size, shape and geometric arrangement. In resolving the reference of objects both kind of attention interact. Furthermore, in dynamic environments as the conversation progresses the attention on objects changes based on object visibility and recency of it being added to the common ground (for discussion see (Kelleher and Dobnik, 2020)). In this paper we examine attention on objects by inspecting how they are referred to in two different tasks using two corpora: the Cups corpus (Dobnik et al., 2020) and the Tell-me-more corpus (Ilinykh et al., 2019).

2 Tasks and corpora

The Cups corpus contains longer English and Swedish dialogues where participants have to identify missing cups on a large table that are hidden to them but these are visible to their conversational partner and vice versa. The cups differ in features such as type, colour and location. Participants are located at the opposite sides of the table and they see each other as an avatar. Figure 1 shows a top-down view of the scene. Each participant sees the same table scene from their own point of view as shown in Figure 7 in Appendix. In addition, there is also a passive observer Katie on the side of the table. Participants are instructed to interact over a chat interface to find the cups each is missing. Beyond this information they are not specifically

Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue, August, 22-24, 2022, Dublin.

told how they should approach the task, the aim is that they negotiate the strategies through their linguistic interaction and engage in a longer dialogue. Table 5 in Appendix shows the overall coverage of the dialogues. The data has been annotated to study different conversational phenomena including spatial perspective taking (Dobnik et al., 2020), dialogue games (Storckenfeldt, 2018; Dobnik and Storckenfeldt, 2018) and reference and coreference (Dobnik and Loáiciga, 2019; Silfversparre, 2021; Dobnik and Silfversparre, 2021). The results reported in this paper are based on these annotations.[1]

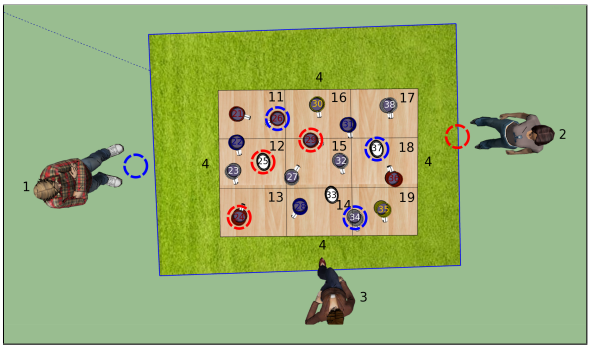


Figure 1: A top-down view of the scene with all objects included and their IDs. Objects marked with coloured circles cannot be seen by a participant marked with the same colour. P3 is a passive observer Katie.

The Tell-me-more corpus (Ilinykh et al., 2019) contains descriptions of images of house environments where participants (via crowd-sourcing) were encouraged to provide multi-sentence descriptions of them. The task can be considered as a simplified form of dialogue with fixed conversational roles of participants. It involves incremental updates of scene descriptions from a describer to an imaginary interactive partner requesting additional information over five turns. The goal of the task is to study incremental referring which is reflected in the discourse structure of the generated text.

We choose these datasets because they provide different scenarios for the study of attention patterns being relevant for the resolution of reference. The Cups scene is known and is identical for all the dialogues. It contains objects of restricted kind, namely the cups, but these vary in terms of their properties such as colour and location. This allows us to study referring over longer sequences of dialogue as well as how participants visually segment larger scenes into smaller regions and how such structuring of a task is reflected in their interaction. Both participants are human, they each have the same goal and by default they do not have pre-determined roles. Instead, these are negotiated between them as the conversation unfolds so that they both can complete the task. The Tell-me-more images are are real-world images different for each discourse where the view of the scene has been determined by the author of the photo. The conversational roles and the view are fixed and consequently interactions are short. However, in this fixed view a variety of scene entities are available that can be potentially referred to. Therefore, the Tell-me-more corpus allows us to study reference and re-reference at a thematic and scene-topological level whereas Cups allows us to study them at the interaction level. Since each involves a different task, a comparison of referring also sheds light on the effect of the task on referring.

While reference in Cups was annotated by human annotators, for Tell-me more we perform this by automatic linking of noun phrases from sequences of image descriptions to object descriptions detected by an object detector. We extract noun phrases from image descriptions using SpaCy (Honnibal et al., 2020). If the head of a noun phrase is not a noun, we consider it an incorrect detection and remove it. We also create a list of words describing types of rooms (e.g., "bedroom", "'attic") based on the (Zhou et al., 2017) hierarchy of images. Overall, we extract 51,953 noun phrases with 9.11 noun phrases per image description and 15,507 noun phrases describing rooms with 2.72 phrases per image description. For object detection we use the model by Anderson et al. (2018).[2] This takes an image and produces a list of detected objects with bounding boxes and object descriptions. The latter include labels (e.g., "chair") and their attributes (e.g., "black"). We limited the number of extracted objects per image to 36.

We explore different methods for linking noun phrases from textual descriptions and object descriptions of detected objects. (Ilinykh and Dobnik, 2022) demonstrate that a transformer-based model and cosine similarity between two phrases (Reimers and Gurevych, 2019) with a threshold 0.5 gives the best performance. For plural noun phrases we follow (Ilinykh et al., 2019) by taking their singular form and link them to objects that have the

[1] https://github.com/sdobnik/cups-corpus

[2] https://github.com/peteanderson80/bottom-up-attention

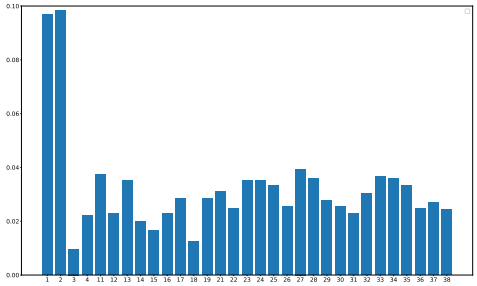
most similar word as head in their description.

We extend the linking method by controlling detection of objects based on the confidence scores of the predicted object labels (L) and attributes (A). We consider three conditions: (i) **AL**, (ii) **[A]L** and (iii) **[A][L]** where brackets indicate that the inclusion of a label or an attribute is conditioned by a confidence score threshold. The thresholds we use are approximations from (Anderson et al., 2018) and were 0.4 for attributes and 0.1 for labels. We evaluate each method manually, by randomly sampling 10 image-text pairs from the dataset and inspecting the correctness of the linking against the expected links, annotated by one of the authors. In 10 image-text pairs there were 102 noun phrases on which each method performed similarly, with the number of incorrect links not exceeding 30. Specifically, **[A][L]** made the fewest errors (25), while **AN** and **[A]N** followed with 28 and 30 errors respectively. One explanation why controlling for both attributes and labels performs best is that it filters out detections with low confidence scores and decreases hallucinations based on textual predictions. As objects with missing labels are removed, it also removes duplicate bounding boxes with low confidence scores. As linking is a highly complex semantic task, no doubt more work is required to improve and evaluate different methods.

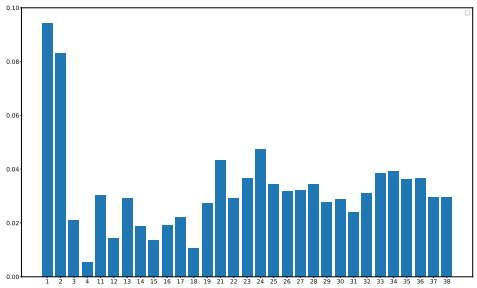
3 Reference in the Cups corpus

Table 1 shows reference to objects in both English and Swedish dialogues. The column *Objects* indicates the number of references to individual objects including both pre-annotated objects (see Figure 1) and objects identified by annotators while the column *Our objects* contains references to pre-annotated objects. Objects identified by annotators include references to parts of objects (e.g. handles of cups and lids) and regions that are created based on the topological arrangement of objects (rather than our pre-annotated regions) (Dobnik and Silfversparre, 2021). In the remainder of the paper we will be referring to these objects. The column *Referring expressions* lists the number of referring expressions in each dialogue. The counts in this column are lower than the counts from the previous column which means that referring expressions contain reference to more than one object, on average two objects per referring expression. However, when one examines the dialogues it can be seen that there is a considerable variation between the number of objects between referring expressions. Some are referring to uniquely identified objects while others are referring to groups of varies sizes. Since dialogues are of different lengths we normalise all three columns to average values per turns. There are differences between individual dialogues but no differences between English and Swedish dialogues. Overall, there are between 3 and 7 objects referred to per turn, when we exclude annotator created objects, between 2 and 5. There are between 1.5 and 3 referring expressions per turn.

3.1 Objects referred to



(a) English dialogues



(b) Swedish dialogues

Figure 2: Reference to entities for English and Swedish dialogues: 1–3 are participants, 4 is the table, 11–19 are regions and 21–38 are objects. To allow comparison all counts are normalised to the total number of references per language corpus, i.e. the columns sum to 1. See also Figure 1 for the representation of the scene objects.

Overall, there are differences in referring to objects between languages. A χ^2 test of independence found a significant relationship between language and reference to scene entities: χ^2(df=30, N=4344)=60.5756, p=0.0008.

Participants (1, 2) most frequently refer to themselves. In the English dialogues the reference to both participants is nearly equal but in the Swedish dialogues participant 1 is more frequently referred to than participant 2. Katie (3), a passive observer

Dialogue	*Length in turns*	*Objects referred to*	*per turn*	*Our objects referred to*	*per turn*	*Referring expressions*	*per turn*
en-1	157	530	3.376	478	3.045	282	1.783
en-2	441	1316	2.984	968	2.195	683	1.549
sv-1	118	407	3.445	261	2.212	177	1.5
sv-2	114	613	5.377	480	4.211	314	2.754
sv-4	75	513	6.84	369	4.92	251	3.347
sv-5	163	628	3.853	473	2.90	334	2.05
sv-6	248	786	3.17	604	2.435	408	1.645
sv-7	308	922	2.994	711	2.309	469	1.523

Table 1: Objects referred and the number of referring expressions in the Cups dialogues.

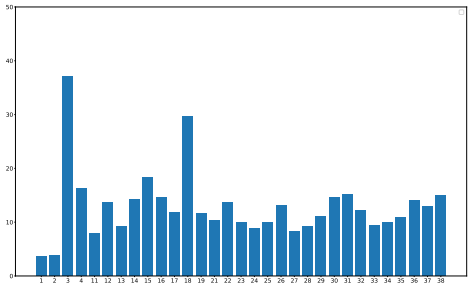
is rarely referred to, even less than objects or regions. This indicates the effect of the task on referring. Participants have a central role in the task (they have to find the missing cups each) as well as they are coordinating the task and the dialogue. Objects and regions are a part of the task. Katie, although animate and therefore potentially a salient landmark, is only a passive observer in this case and does not contribute to the task. The table (4) is more frequently referred to in the English than Swedish dialogues but overall it is among less frequently referred to entities, possibly serving as a landmark in descriptions involving top-view allocentric frame of reference. The next type of entities ranked by the increasing frequency are regions (11–19). Here we see that in both groups of dialogues regions 11, 13, 17 and 19 are most frequently referred to whereas region 18 is the least frequently referred to region. Figure 1 shows that these are the corner regions of the table, hence regions of the table that are closest to each participant and on their left and right. Regions that are between these regions receive less attention, most notably region 18 which is the central region closer to P2. Overall, objects are even more frequently referred to than regions. Here there is a slight difference between languages for example some most frequently objects referred to in English are 27, 28, 33, 34, 23, 24 and for Swedish 24, 21, 33, 34, 35, 36, 28. Examining the scene in Figure 1 we can see that are related to the missing cups 24, 25, 26, 29, 34, 37 either because they are the missing cup (e.g. 24, 34), they are a distractor object for the missing cup (i.e. the cup that could be referred to with the same description as the missing cup, e.g. 21 for 26 or 24, and 33 and 35 for 34.) For example, 28, on the other hand, is a cup proximal to two missing cups and therefore a good landmark to refer to to resolve the task. A considerable part of the dialogue involves resolving reference of these descriptions and there are sections of dialogue where a describer and interpreter (who later also becomes a describer) refer to different entities with the same description until a contradiction is detected and diverged commons grounds are reconciled (for example, en-1 turns 42–62). Object 31 is the least frequently referred to in dialogues of both languages. This is a blue cup close to the missing red cup 29 and the white cup 37. As such it is not a distractor object to either of them and therefore likely be used only as a landmark for reference to other cups in configurations where other landmarks are also possible candidates (the same holds for object 22). Overall, the results indicate that the task and the way the scene was structured through the introduction of the missing cups has an effect on the attention the objects receive through reference.

The proportion of objects therefore tells us about their perceptual and task related salience. Frequently referred to objects are those that are related to the task but also those that are perceptually salient either because they are visually similar to the target objects or because they are good landmarks that target objects can be described with, for example the corner regions of the table. It is also observable that both properties interact. For example, visually accessible regions on the lateral dimension of the scene are more perceptually accessible to participants than the front and back regions and therefore they are more frequently referred to.

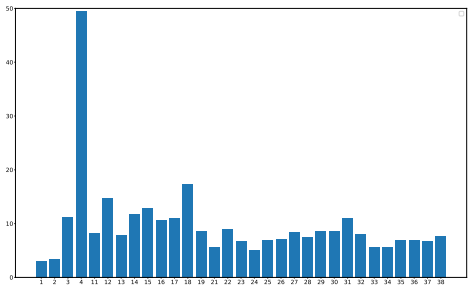
	en-1	en-2	sv-1	sv-2	sv-4	sv-5	sv-6	sv-7
en-1		ns	***	***	ns	ns	ns	**
en-2	ns		***	***	*	*	ns	***
sv-1	***	***		***	***	***	***	***
sv-2	***	***	***		ns	ns	ns	***
sv-4	ns	*	***	ns		ns	**	*
sv-5	ns	*	***	ns	ns		ns	***
sv-6	ns	ns	***	ns	**	ns		***
sv-7	**	***	***	***	*	***	***	

Table 2: χ^2 test of independence comparing reference to scene entities across dialogues. *** indicates $p < 0.001$, ** $p < 0.01$, * $p < 0.05$ and ns indicates non-significant difference. For details see Table 6 in Appendix.

What is the variation of reference to objects between different conversational participant pairs in both languages? Table 2 shows the results of a χ^2 test of whether reference to scene entities differs between different dialogues. For English we only have two dialogues, en-1 and en2, for which the test indicates that they do not differ in reference to objects. The en-1 dialogue is more similar to the Swedish dialogues than en-2 which is an interesting observation since the speakers of en-1 are native speakers of Swedish (see Table 5 in Appendix). This suggest that there may be differences between languages in terms of referring. Among the Swedish dialogues sv-1 and sv-7 stand out as the reference there differs to reference in all other dialogues. The reference in the remaining Swedish dialogues, sv-2, sv-4, sv-5 and sv-6 is similar, except for dialogues sv-4 and sv-6 where reference is different but only when compared with each other. The results point that although different pairs of conversational participants structure the task freely and sometimes very differently, objects are still referred to in the same way. There is also an effect of language and possibly the way conversational participants approached to solve the task and their individual preferences.



(a) English dialogues



(b) Swedish dialogues

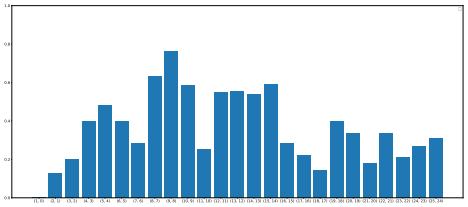
Figure 3: Mean distance between turns that repeat reference to entity for English and Swedish dialogues: 1–3 are participants, 4 is the table, 11–19 are regions and 21–38 are objects. See also Figure 1.

3.2 Re-reference to objects

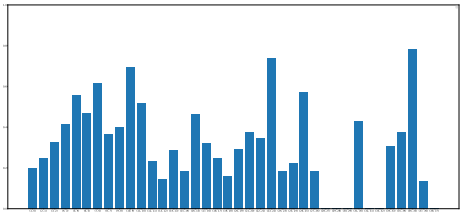
How likely is that an object will be re-referred in subsequent turns? Figure 3 shows for each scene entity the average distance (separation) between turns when this entity is re-referred in both English and Swedish dialogues. We estimate distance between each consecutive pair of turns when a particular entity has been referred to. Re-reference shows similar trends for both English and Swedish dialogues for individual scene entities. However, overall, the distance between turns over which they are re-referred is slightly greater in English than Swedish. This excludes object 4, the table. As expected, large distance of re-reference is associated with low frequency. Participants 1 and 2 are re-referred most recently but also most frequently (see Figure 2). Similarly, objects 24, 27, 28, 34, 34 for English and 24, 21, 34, 34 for Swedish. Katie (3) and table (4) are re-referred to a greater number of turns apart but also very infrequently. Similarly, objects 18, 15, 30, 31 for English and 18, 12, 15, 31 for Swedish. Overall, regions are re-referred after a greater number of turns than objects in both English and Swedish dialogues. Regions 11 and 13 are the most recently re-referred regions both in English and Swedish which is again associated with their high frequency. That regions are re-referred after greater number of turns than objects again confirms that they serve as landmarks for identifying objects when needed while objects are the main targets of descriptions identified by the task.

A non-uniform distribution in which objects and regions are re-referred indicates that these are not referred to randomly as the dialogue progresses. Work on dialogue interaction (Clark, 1996) and as well as previous work on the Cups conversations indicate that participants split the task, the scene and therefore conversations into sub-parts. The Swedish dialogues have been annotated for dialogue games (Kowtko et al., 1992; Carletta et al., 1997) with two kinds of tags, one indicating the scope of the games over turns and one indicating the type of the games (Storckenfeldt, 2018). Dialogue games can be nested, a typical example being a clarification game which is embedded in another game. In the next experiment we measure to what degree objects referred to in one dialogue game overlap with the objects referred to in other

dialogue games. As a measure of overlap we use Sørensen–Dice coefficient $DSC = \frac{2|A \cap B|}{|A|+|B|}$ which ranges between 0 (no overlap) and 1 (perfect overlap). Note that here we calculate overlap of sets which means that duplicate reference is counted only once. As individual conversations structured differently in terms of dialogues games and strategies to refer to objects we represent these for each dialogue game separately. When comparing pairs of adjacent games for objects they are referring to we obtain the mean values of DSC and their standard deviations as follows: sv-1: μ=60.7 σ=24.1, sv-2: μ=36.2 σ=18.6, sv-4: μ=31.7 σ=21.1, sv-5: μ=31 σ=21.2, sv-6: μ=37 σ=25.1 and sv-7: μ=29.5 σ=23.9. The results indicate that except for sv-1 where there is a high overlap of objects referred to across adjacent games (60.7), adjacent dialogue games overlap in reference in about a 1/3 according to DSC. However, notably there is a high standard deviation which indicates a high variability between individual games.



(a) sv-2



(b) sv-5

Figure 4: Dice-Sørensen coefficient of reference overlap between adjacent dialogue games.

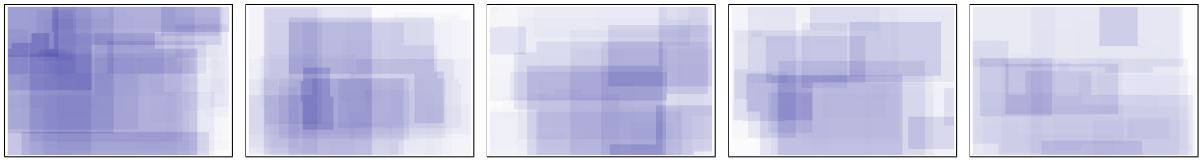
Figure 4 shows a comparison of two dialogues from which it can be seen that referring overlap is non-uniform and there are sections of dialogue where there is either a high or a low degree of reference overlap between adjacent games. In Figure 8 in Appendix we cross-tabulate DSC for all games (i.e. not just adjacent ones). The plots indicate that reference is local and restricted to a couple of adjacent dialogue games but a reference to the same objects might be made at a later stages of dialogue, again with a local scope. Overall, this indicates that reference to objects is highly dependent on how conversational partners negotiate and structure their task. Conversationally, structuring a large scene into local sub-parts has a referring advantage as expressions can be made more optimal and be less ambiguous (for example, by requiring less descriptive attributes) as attention is placed on a smaller number of distractor objects that are potential referents.

4 Reference in the Tell-me-more corpus

4.1 The location of objects referred to

Where (Landau and Jackendoff, 1993; Landau, 2016) in the image frame are these objects located? We track attention to objects in images by representing the overlap of the bounding boxes of objects referred to in each of the five (5) sentences that constitute a single image description. To demonstrate the effects of the discourse we represent attention maps collectively for all images for the first, second, ..., fifth sentence of the discourse. First, we take all images and re-scale them to $T \times T$ pixels, where $T = 2000$. Along with the images, we also resize bounding boxes accordingly to ensure that they correspond to the detected objects in size and location. For each sentence in a sequence, we draw a heat-map from bounding boxes of those objects that are mentioned in that sentence. In order to generate a single heat-map per sentence across all images, we use alpha blending (Blinn, 1994), a method that takes an image and maps another image on top of it. The mapping is controlled by two α values which determine the transparency of each of the two images. We set them to 0.9 and 0.1 for the background and foreground mapped images respectively. We also normalise the resulting heat-maps by the number of images in the dataset.

Figure 5 shows five attention heat-maps with darker areas indicating attention to objects being referred to. In general, the first sentence refers to the most of the scene, there is a high overlap between object bounding boxes, changing their attention on specific parts and objects later in the sequence. This finding shows the sequential nature of image description sequences and aligns with the idea that humans structure scene discourse and mention objects in some order (Grosz and Sidner, 1986). Note that there is also an impact of the type of the visual



(a) Sentence 1 (b) Sentence 2 (c) Sentence 3 (d) Sentence 4 (e) Sentence 5

Figure 5: Attention maps of bounding boxes of objects referred to determined through automatic linking. The results are averaged per sentences and across all images and discourses.

scenes as humans tend to describe apartment layouts through the "tour strategy", organising spatial descriptions of the house environments in a tree structure (Linde and Goguen, 1980).

We note that among sentences 2, 3 and 4 the attention shifts from one side of the image to another on the later dimension. This could be related to the fact that left-right is a prominent relation used in spatial inference along which target and landmark objects are related. In the last sentence the attention is generally weaker, indicating much fewer and smaller objects described. The number of objects linked on average per image is 3.94, 2.38, 2.02, 1.79, 1.60 for sentence from 1 to 5 respectively, showing that humans start with detailed descriptions of images and later focus on smaller parts of the scene, describing fewer objects. Overall, the results indicate that the attention on the image changes over the discourse: from several larger objects to fewer and smaller objects. There is also evidence of spatial inference in the left-right axis.

4.2 Thematic associations between objects

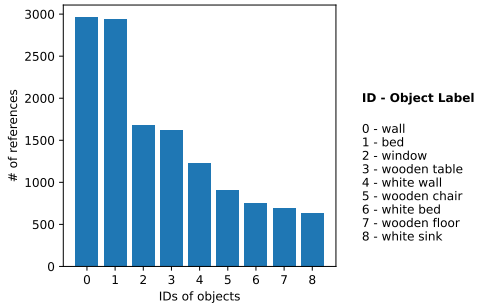


Figure 6: The frequency of top-*k* referred to objects across all images as determined by automatic linking.

What are the most frequently described objects across all image description sequences in the Tell-me-more corpus? Figure 6 shows the top frequencies of objects referred to as determined by automatic linking. These objects are typically objects that occur across many different room types: walls, chairs, floors, tables. Among 1,686 described object types, 560 of them (approximately 33%) are described only once. These least frequently described objects include, for example, "gray skirt", "orange stripe" and "beige light switch".

To investigate how thematic relations between objects drive attention as expressed through reference we examine and compare (i) how objects appear in situational contexts and (ii) how these objects are referred to in descriptive discourses. For this we build *vector space models* (Turney et al., 2010) of object co-occurrences where context counts are either determined by (i) scene co-occurrence and (ii) scene and referential co-occurrence. The *scene vector space* captures the number of times each object appears in the scene with other objects. We consider objects which have acceptable confidence scores for both attributes and labels as determined by the **[A][N]** linking method. Table 8 in Appendix shows a part of the scene co-occurrence vector space. The size of this space is $3,054 \times 3,054$. Note that this is greater than the number of objects labels in the model $(1,600)$, because object descriptions also contain attributes which introduces permutations of descriptions. The *referential vector space* captures for each object the frequency of it being mentioned together with other referentially linked objects. Table 7 in Appendix shows an excerpt from the constructed referential vector space. The size of the space is 1686×1686 which is smaller than the size of the scene vector space. This indicates that a large number (44.8%) of scene objects is not mentioned at all (including false negatives of our automatic linking method).

What are the differences between the two vector spaces? To examine the relationship between frequencies of objects in two vector spaces we compute a Spearman's rank correlation coefficient. We observe a positive correlation between the two variables, $r = 0.82$, $p = .000$, demonstrating that the frequencies of objects in images correlate with the frequencies of them being mentioned in image de-

scriptions, subject to the accuracy of linking. This is of no surprise since if an object is in an image, it has a certain probability to be mentioned in the image description.

For each vector space we rank the objects by their frequency of occurrence and then extract their ten most similar objects using the kd-tree (Maneewongvatana and Mount, 1999) which is an improved version of the k-nearest neighbour algorithm. Figure 9 and Figure 10 in Appendix show the most frequently and the least frequently occuring objects in both vector spaces and 10 of their most similar objects. The results indicate that the most frequently occuring objects are similar for both configurations. They include objects most commonly found in rooms such as wall, window, table and chair. However, there is a difference in what their most similar objects are. It appears that the similarity of objects from the referential vector space is based on the attributes and not just object co-occurrence, e.g. wooden table: wooden floor, white wall: white lamp, white window. This indicates that semantic distinctions captured are not only based on situational co-existence but other dimensions of meaning defined by the attribute: i.e. objects of the same colour or consisting of the same material (cf. the semantic distinction between sense and reference). For the least frequently occurring objects there is a high variation both in terms of what these objects are and their most similar objects in the corresponding vector spaces. This is expected because of their low frequency support.

Table 3 shows three objects and their most similar objects in both vector spaces. The referential vector space captures also *thematic* relations between objects: "stainless steel oven" is similar to "blender" and "silver coffee maker" which fit into a thematic cluster of kitchen appliances. On the other hand, the scene vector space captures similarities of co-occuring objects: it predicts "brown pot" and "white floor" similar to "stainless steel oven". Referential vector space therefore also encodes information about how humans group objects in scenes and describe them within a depiction of same event or a task. Other words show similar trends: "marble counter" is similar to bowls, knobs, food, bananas and hair dryers indicating other objects that interact with marble counters. On the other hand shelves, windows, refrigerators and ceiling predicted by the scene vector space are co-occuring objects in the same rooms. This shows that the task and subsequent human communicative intents are important factors of what gets included in a description: objects are not only described because they are there, but because they are thematically connected with each other at a higher task-related level.

4.3 Attention to objects through reference

Can we estimate this thematic attention for the objects referred to in the Tell-me more dataset? From objects appearing in a scene, what objects (i) are likely to be referred to, (ii) are likely to be re-referred in the same discourse, and (iii) are likely not to be referred to? To answer these questions, for each object w_n we compute *attention* as a ratio A_{w_n} between its vector in the reference vector space $\mathbf{V}^r$ and the scene vector space $\mathbf{V}^s$:

$$A_{w_n} = \frac{\sum \mathbf{v}^r_{w_n}}{\sum \mathbf{v}^s_{w_n}}, \quad (1)$$

where $\mathbf{v}^*_{w_n}$ is a word frequency vector in the corresponding vector space. An attention score 1 indicates that an object is referred every time when it occurs in an image. An attention score > 1 indicates that an object is likely to be re-referred in the same discourse and an attention score < 1 indicates that an object is referred to less frequently than it occurs. Attention scores close to 0 indicate that objects are nearly never referred to. Therefore, the resulting attention scores can be interpreted as *thematic salience* of objects in this domain.

Table 4 shows some of the most and the least attended objects in this corpus. First we note that 1,368 out of 3,054 objects are assigned an attention score 0.0 because they are never referred to (subject to the automatic linking method). Object names of the most attended objects often include attributes which refer to colour (e.g., "green stripe", "white artwork"). For example, "painted wall" is likely to be referred to (attention score 2.375) but "wall" has a score of 0.210848. While "black horse" is highly attended (ranked 13 among 1,686 objects), "black faucet" is ranked 1,672. This could be an artefact of using phrase similarity to match descriptions with object names containing attributes. It could be that the colour of the faucets is less likely to be described than the colour of horses and therefore an object label "black faucet" is less likely to be matched with a description "faucet". Similarly, "orange flower" and "white freezer" are unlikely to be referred to with these attributes while "blue flowers" are more likely with an attention score of

stainless steel oven		**silver refrigerator**		**marble counter**	
Ref space	*Scene space*	*Ref Space*	*Scene space*	*Ref Space*	*Scene space*
stainless steel oven	stainless steel oven	silver refrigerator	silver refrigerator	marble counter	marble counter
backsplash	white stove	silver microwave	white rug	black handle	mantle
tiled backsplash	red fruit	stainless steel dishwasher	marble counter	white bowl	shelf
white backsplash	yellow bottle	silver stove	clear wine glass	white hair dryer	brown window
white blender	brown pot	stainless steel oven	bowl	white knob	silver refrigerator
metal hood	white floor	stainless steel refrigerator	brown floor	green bananas	hanging chandelier
silver coffee maker	white table	black microwave	round table	gray towel	ceiling
pink bottle	pink cushion	oven	black printer	food	wooden floor
clear wine glass	white windows	brown cabinets	mantle	black light	rug
silver dishes	hanging light	food	shelf	stainless steel stove	black table
white lights	wooden wall	black oven	wooden chair	cabinets	white lamp

Table 3: The most similar objects for three target objects in referential and scene vector spaces. Objects are ordered from most (top) to least similar (bottom).

	Object	Attention score
0	green stripe	3.428571
1	white artwork	3.375000
2	red comforter	2.888889
3	decorative painting	2.823529
4	colorful couch	2.500000
5	painted wall	2.375000
6	white chicken	2.187500
7	seat	2.166667
8	yellow game	2.000000
9	black barrel	1.993243
10	pink sink	1.928571
11	silver drawers	1.800000
12	black horse	1.722222
13	gold headboard	1.684211
14	gold ceiling	1.666667
15	brown horse	1.664000
16	purple table	1.652174
17	leather recliner	1.642857
18	black machine	1.421687
19	white clothes	1.411765
...	...	...

	Object	Attention score
...	...	...
1666	hanging chain	0.015748
1667	red room	0.015504
1668	red shelf	0.014925
1669	white freezer	0.014787
1670	red rack	0.014706
1671	orange flower	0.014184
1672	white cup	0.013514
1673	yellow bottle	0.012121
1674	wooden entertainment center	0.011792
1675	black tire	0.011164
1676	red door	0.010870
1677	pot	0.010063
1678	black faucet	0.009740
1679	handle	0.008611
1680	outlet	0.007282
1681	yellow bowl	0.007067
1682	vent	0.006589
1683	parked car	0.006494
1684	metal pole	0.005952
1685	silver shower head	0.004907

Table 4: Attention scores for twenty most attended (left) and least attended objects (right).

0.524193. The attribute salience described here is common-sense thematic salience which is different from visual salience (Kelleher et al., 2005). It is important to note that both kinds of salience interact. For example, a "white freezer" is more likely to be referred to in the context of all black freezers.

5 Discussion and conclusion

Our comparison of reference in the Cups and Tell-me-more corpora reveals several factors that affect what objects are referred to and when. Referring is highly influenced by the nature of the conversational tasks which shapes the goals of participants and is reflected in conversational interaction. Participants in the Cups dialogues have identical conversational roles and are free to structure their interactions. On the other hand the task of referring in the Tell-me-more corpus and the roles of participants is highly restricted but so are the patterns of reference produced. Furthermore, we can observe differences in referring to scene entities as the discourse progresses. Therefore, it is *wrong* to assume that Tell-me-more and image captioning in general represent a task-neutral setting. Previously, referring expressions have been studied only within a particular corpus or a task but our findings indicate that this is by no means sufficient to understand referring. Further examination of the task structure which is reflected in discourse, for example in conversational games, might point to common referring patterns between tasks and make the notion of the task less elusive. We have also identified other factors relevant for referring: visual properties of the scene, geometric arrangements of scene objects and patterns of spatial reasoning. There are thematic relations between objects that go beyond the presence of objects in the scene and are related to description of coherent events.

Referring is a complex phenomenon that is hard for computational modelling. As it is context and task dependent this means that large corpora will have to be available to capture all the tasks, that involve referring. Focusing on simple tasks such as image captioning or dialogues with a single dialogue game is not enough. The task dependence has implications for transfer learning as this should be difficult between tasks that differ considerably. This could be the reason why using language-independent object detection in multi-modal NLP tasks with language-based transformers is better than utilising pre-trained visual embeddings which have been trained together with language. This way an interaction model can be trained separately and specifically for each task.

Acknowledgements

The research reported in this paper was 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

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6077–6086.

J.F. Blinn. 1994. Compositing. 1. theory. *IEEE Computer Graphics and Applications*, 14(5):83–87.

Donna K Byron. 2003. Understanding referring expressions in situated language some challenges for real-world agents. In *Proceedings of the First International Workshop on Language Understanding and Agents for Real World Interaction*, pages 39–47.

Jean Carletta, Stephen Isard, Gwyneth Doherty-Sneddon, Amy Isard, Jacqueline C Kowtko, and Anne H Anderson. 1997. The reliability of a dialogue structure coding scheme. *Computational linguistics*, 23(1):13–31.

Herbert H. Clark. 1996. *Using language*. Cambridge University Press, Cambridge.

Kees van Deemter. 2016. *Computational models of referring: a study in cognitive science*. The MIT Press, Cambridge, Massachusetts and London, England.

Simon Dobnik, John D. Kelleher, and Christine Howes. 2020. Local alignment of frame of reference assignment in English and Swedish dialogue. In *Spatial Cognition XII: Proceedings of the 12th International Conference, Spatial Cognition 2020, Riga, Latvia*, pages 251–267, Cham, Switzerland. Springer International Publishing.

Simon Dobnik and Sharid Loáiciga. 2019. On visual coreference chains resolution. In *Proceedings of LondonLogue – Semdial 2019: The 23rd Workshop on the Semantics and Pragmatics of Dialogue*, pages 1–3, London, UK. Queen Mary University of London.

Simon Dobnik and Vera Silfversparre. 2021. The red cup on the left: Reference, coreference and attention in visual dialogue. In *Proceedings of PotsDial - Semdial 2021: The 25th Workshop on the Semantics and Pragmatics of Dialogue*, Proceedings (SemDial), pages 50–60, Potsdam, Germany.

Simon Dobnik and Axel Storckenfeldt. 2018. Categorisation of conversational games in free dialogue over spatial scenes. In *Proceedings of AixDial – Semdial 2018: The 22st Workshop on the Semantics and Pragmatics of Dialogue*, pages 1–3, Aix-en-Provence, France.

Barbara J Grosz and Candace L Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational linguistics*, 12(3):175–204.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Nikolai Ilinykh and Simon Dobnik. 2020. When an image tells a story: The role of visual and semantic information for generating paragraph descriptions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 338–348, Dublin, Ireland. Association for Computational Linguistics.

Nikolai Ilinykh and Simon Dobnik. 2022. Attention as grounding: Exploring textual and cross-modal attention on entities and relations in language-and-vision transformer. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4062–4073, Dublin, Ireland. Association for Computational Linguistics.

Nikolai Ilinykh, Sina Zarrieß, and David Schlangen. 2019. Tell me more: A dataset of visual scene description sequences. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 152–157, Tokyo, Japan. Association for Computational Linguistics.

John D. Kelleher, Fintan J. Costello, and Josef van Genabith. 2005. Dynamically structuring updating and interrelating representations of visual and linguistic discourse. *Artificial Intelligence*, 167(1):62–102.

John D. Kelleher and Simon Dobnik. 2020. Referring to the recently seen: reference and perceptual memory in situated dialogue. In *CLASP Papers in Computational Linguistics: Dialogue and Perception – Extended papers from DaP-2018 Gothenburg*, volume 2, pages 41–50, Gothenburg, Sweden. University of Gothenburg, CLASP, Centre for Language and Studies in Probability and GUPEA.

Jacqueline C Kowtko, Stephen D Isard, and Gwyneth M Doherty. 1992. Conversational games within dialogue. HCRC research paper RP-31, University of Edinburgh.

Barbara Landau. 2016. Update on "what" and "where" in spatial language: A new division of labor for spatial terms. *Cognitive Science*, 41(2):321–350.

Barbara Landau and Ray Jackendoff. 1993. "What" and "where" in spatial language and spatial cognition. *Behavioral and Brain Sciences*, 16(2):217–238, 255–265.

Charlotte Linde and J.A. Goguen. 1980. On the independence of discourse structure and semantic domain.

In *18th Annual Meeting of the Association for Computational Linguistics*, pages 35–37, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Sharid Loáiciga, Simon Dobnik, and David Schlangen. 2021. Annotating anaphoric phenomena in situated dialogue. In *Proceedings of the 1st Workshop on Multimodal Semantic Representations (MMSR) at IWCS 2021*, pages 78–88, Groningen, Netherlands (Online). Association for Computational Linguistics.

Didier Maillat. 2003. *The semantics and pragmatics of directionals: a case study in English and French*. Ph.D. thesis, University of Oxford: Committee for Comparative Philology and General Linguistics, Oxford, United Kingdom.

Songrit Maneewongvatana and David M. Mount. 1999. Analysis of approximate nearest neighbor searching with clustered point sets. In *Data Structures, Near Neighbor Searches, and Methodology*.

Massimo Poesio, Yulia Grishina, Varada Kolhatkar, Nafise Moosavi, Ina Roesiger, Adam Roussel, Fabian Simonjetz, Alexandra Uma, Olga Uryupina, Juntao Yu, and Heike Zinsmeister. 2018. Anaphora resolution with the ARRAU corpus. In *Proceedings of the First Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 11–22, New Orleans, Louisiana. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Vera Silfversparre. 2021. The red cup on your left: Reference, coreference and visual attention in visual dialogue. C-uppsats (bachelor's thesis/extended essay), Department of Philosophy, Linguistics and Theory of Science (FLOV), University of Gothenburg, Gothenburg, Sweden. Supervisor: Simon Dobnik, examiner: Moa Ekbom.

Manfred Stede. 2011. Discourse processing. *Synthesis Lectures on Human Language Technologies*, 4(3):1–165.

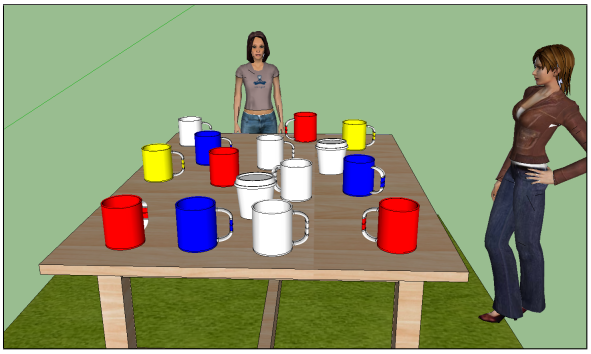
Axel Storckenfeldt. 2018. Categorisation of conversational games in free dialogue referring to spatial scenes. C-uppsats (bachelor's thesis/extended essay), Department of Philosophy, Linguistics and Theory of Science (FLOV), University of Gothenburg, Gothenburg, Sweden. Supervisor: Simon Dobnik, examiner: Ylva Byrman.

Ece Takmaz, Mario Giulianelli, Sandro Pezzelle, Arabella Sinclair, and Raquel Fernández. 2020. Refer, Reuse, Reduce: Generating Subsequent References in Visual and Conversational Contexts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4350–4368, Online. Association for Computational Linguistics.

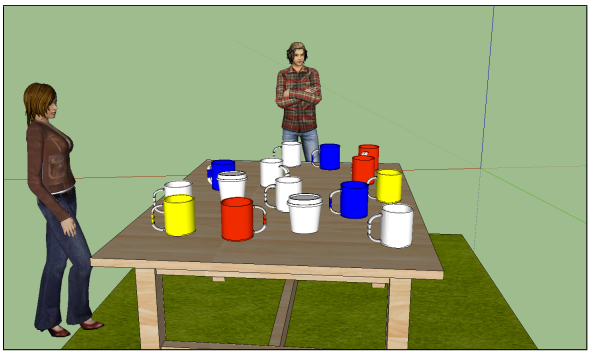
Peter D Turney, Patrick Pantel, et al. 2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37(1):141–188.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. 2017. Scene parsing through ade20k dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5122–5130.

A Appendix



(a) The view of P1

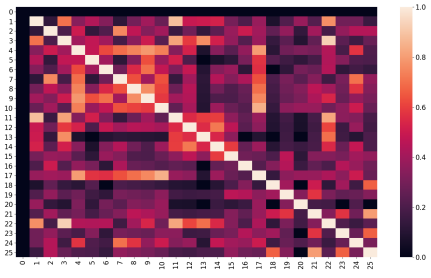


(b) The view of P2

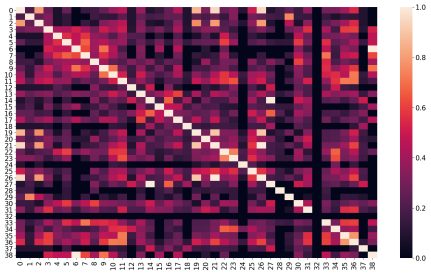
Figure 7: The scene as seen by P1 (a) and P2 (b). P3 is a passive observer Katie.

Dlg	*Language*	*Native*	*Duration (min)*	*Length (turns)*
en-1	English	Swedish	≈30	157
en-2	English	British	≈60	441
sv-1	Swedish	Swedish	≈80	118
sv-2	Swedish	Swedish	≈40	114
sv-4	Swedish	Swedish	≈30	75
sv-5	Swedish	Swedish	≈60	163
sv-6	Swedish	Swedish	≈60	248
sv-7	Swedish	Swedish	≈60	308

Table 5: The coverage of the Cups corpus per dialogues



(a) sv-2



(b) sv-5

Figure 8: Dice-Sørensen coefficient of reference overlap for all pairs of dialogue games.

Dialogue	χ^2	*N*	*p*	*dof*	sig
sv* vs en*	60.5756	4344	0.0008	30	***
en-1 vs en-2	29.0450	1446	0.5152	30	ns
en-1 vs sv-1	86.7180	739	2.05E-07	30	***
en-1 vs sv-2	71.0177	958	3.54E-05	30	***
en-1 vs sv-4	26.4639	847	0.6513	30	ns
en-1 vs sv-5	25.7953	951	0.6855	30	ns
en-1 vs sv-6	36.5352	1082	0.1912	30	ns
en-1 vs sv-7	57.3736	1189	0.0019	30	**
en-2 vs sv-1	130.6225	1229	1.61E-14	30	***
en-2 vs sv-2	80.8790	1448	1.48E-06	30	***
en-2 vs sv-4	48.0981	1337	0.0194	30	*
en-2 vs sv-5	47.6964	1441	0.0212	30	*
en-2 vs sv-6	35.2543	1572	0.2335	30	ns
en-2 vs sv-7	83.5087	1679	6.11E-07	30	***
sv-1 vs sv-2	99.9123	741	1.92E-09	30	***
sv-1 vs sv-4	89.3357	630	8.28E-08	30	***
sv-1 vs sv-5	130.1799	734	1.92E-14	30	***
sv-1 vs sv-6	117.1598	865	3.04E-12	30	***
sv-1 vs sv-7	84.5977	972	4.22E-07	30	***
sv-2 vs sv-4	32.3097	849	0.3533	30	ns
sv-2 vs sv-5	37.9412	953	0.1513	30	ns
sv-2 vs sv-6	38.9441	1084	0.1270	30	ns
sv-2 vs sv-7	70.8348	1191	375E-05	30	***
sv-4 vs sv-5	30.1364	842	0.4587	30	ns
sv-4 vs sv-6	52.2275	973	0.0072	30	**
sv-4 vs sv-7	48.9344	1080	0.01597	30	*
sv-5 vs sv-6	41.4435	1077	0.0798	30	ns
sv-5 vs sv-7	84.6866	1184	4.10E-07	30	***
sv-6 vs sv-7	83.4423	1315	6.25E-07	30	***

Table 6: χ^2 test of independence comparing reference to scene entities across different dialogues. *** indicates $p < 0.001$, ** $p < 0.01$, * $p < 0.05$ and ns indicates no statistically significant association.

	bed	lit lamp	white ceiling	framed picture	room	wooden nightstand	yellow wall	white rug	···
bed	9030	57	239	593	256	209	166	0	···
lit lamp	57	6	5	6	5	5	6	0	···
white ceiling	239	5	206	66	67	3	40	3	···
framed picture	593	6	66	1144	58	8	29	1	···
room	256	5	67	58	278	7	34	0	···
···	···	···	···	···	···	···	···	···	···

Table 7: A sample of the referential vector space. The values represent the frequencies of objects being referred to together in the same discourse. For example, the object "bed" is referred to in the same discourse with the object "white ceiling" 239 times.

	wall	bed	wooden headboard	white wall	lit lamp	white ceiling	pillow	framed picture	
wall	8262	3167	4916	3458	3385	3991	2685	2207	···
bed	3167	7138	3074	2500	2961	2278	1657	1843	···
wooden headboard	4916	3074	6154	2482	2504	1686	2111	1699	···
white wall	3458	2500	2482	3554	2412	2137	3029	1199	···
lit lamp	3385	2961	2504	2412	1326	1930	1536	1442	···
···	···	···	···	···	···	···	···	···	···

Table 8: A sample of the situation vector space. The values represent the frequencies of objects occuring together with other objects in scenes. The rows and columns do not directly match the rows and columns in the referential vector space because not all objects are referred to in descriptions.

	Most similar objects (ranked from 0 to 9)								
	wall	bed	window	wooden table	white wall	wooden chair	white bed	wooden floor	white sink
0	wall	bed	window	wooden table	white wall	wooden chair	white bed	wooden floor	white sink
1	window	wall	white wall	window	white lamp	wooden table	white pillow	white ceiling	bathroom
2	wooden table	window	wooden floor	wooden floor	white window	wall	white lamp	room	mirror
3	framed picture	white pillow	lamp	white wall	lamp	window	white window	white window	white toilet
4	white wall	white wall	glass window	brown chair	room	wooden floor	white pillows	green plant	white bathtub
5	lamp	white lamp	white lamp	white ceiling	floor	white wall	room	light	white towel
6	wooden floor	lamp	floor	floor	white ceiling	black chair	white wall	floor	large mirror
7	white pillow	framed picture	framed picture	glass window	wooden floor	brown chair	brown pillow	large window	tiled floor
8	white lamp	white window	white ceiling	framed picture	wooden headboard	chair	white shade	ceiling	white tub
9	floor	wooden headboard	white window	lamp	white shade	glass window	lamp	white door	brown floor
	Least similar objects (ranked from 0 to 9)								
	patio	black hair dryer	blue drawers	blue plant	red plant	black windows	off television	wooden cross	brown ground
0	patio	black hair dryer	blue drawers	blue plant	red plant	black windows	off television	wooden cross	brown ground
1	large curtains	wooden doors	store	store	black toaster	glass dish	green bedspread	black shadow	red rack
2	gold lights	green bathroom	brown staircase	white dishes	open book	hanging mirror	blue recliner	hanging chain	wooden doors
3	open doorway	yellow frame	bird	yellow machine	blue recliner	toothbrush	yellow frame	black mantle	yellow frame
4	black pipe	bird	yellow frame	red balloon	wine glass	robe	green fence	orange light	green bathroom
5	gold rod	brown staircase	wooden doors	red kettle	yellow kite	gold hook	wooden doors	small toy	green fence
6	glass pitcher	green fence	red head	gold light switch	metal towel rack	white bucket	green bathroom	wooden light	bird
7	wooden cross	green bedspread	green bathroom	white star	beige sofa	dark window	brown staircase	black hair dryer	brown staircase
8	green drawer	yellow sink	green fence	red head	blue comforter	black holder	bird	wooden plate	pink basket
9	small toy	brick floor	yellow sink	white robe	plaid towel	orange soap	gray door	gray telephone	yellow sink

Table 9: Column names indicate either the most frequently occurring objects in images (the top section of the table) or the least frequently occurring ones (the bottom section of the table). Similarities are calculated on the **referential vector space** and objects are ordered from the most similar (0) to the least similar (9).

	Most similar objects (ranked from 0 to 9)								
0	wall	white wall	window	wooden table	room	framed picture	wooden chair	white ceiling	white window
1	wooden headboard	lit lamp	white wall	lit lamp	white wall	lit lamp	white wall	room	framed picture
2	white wall	white wall	lit lamp	pillow	white ceiling	white wall	lit lamp	wooden nightstand	wooden nightstand
3	bed	wooden headboard	wall	white ceiling	framed picture	framed picture	framed picture	white rug	white rug
4	white ceiling	white ceiling	bed	framed picture	room	wooden table	wooden nightstand	tiled floor	wooden table
5	lit lamp	room	white ceiling	room	tiled floor	tiled floor	room	yellow wall	tiled floor
6	pillow	framed picture	pillow	wooden headboard	wooden table	room	yellow wall	wooden table	green plant
7	framed picture	wall	framed picture	wooden nightstand	wooden nightstand	green plant	white rug	round table	white vent
8	room	pillow	room	tiled floor	green plant	white rug	round table	silver refrigerator	silver refrigerator
9	yellow wall	wooden table	wooden nightstand	wooden table	pillow	wooden nightstand	tiled floor	white vent	black printer
10	wooden nightstand	wooden nightstand	yellow wall	white rug	white rug	white vent	silver refrigerator	green plant	clear wine glass
	Least similar objects (ranked from 0 to 9)								
	gray skirt	black handlebars	black hair dryer	large statues	beige light switch	small shelf	pole	purple cabinet	orange stripe
0	purple stripe	stone column	silver toilet brush	brown umbrella	red booth	blade	blue stairway	striped floor	small sailboat
1	blue tablecloth	brown soap	red door frame	large statues	wooden bowl	pink shoes	pink shoes	glass refrigerator	brown soap
2	red boat	girl	egg	gold candle	blue mouse pad	long tie	glass shower door	metal chain	pink shoes
3	mountains	pink shoes	dark nightstand	blue mouse pad	pink shoes	purple table	orange counter	blue mouse pad	black bird
4	gold tree	blue mouse pad	palm plant	brown soap	blue doors	purple bird	blue mouse pad	pink shoes	girl
5	blue mouse pad	blue backsplash	gray skirt	girl	orange counter	black handles	wooden pen	decorative wall	black handles
6	metal chain	metal stand	blade	pink shoes	purple shade	white hole	black button	metal lock	silver heater
7	bucket	wooden pen	pink shoes	decorative wall	glass shower door	black person	decorative wall	circular mirror	blue mouse pad
8	blue radiator	decorative wall	trees	black button	blue keyboard	candles	girl	brown soap	plastic chair
9	silver wheel	black handles	gold light switch	tan surfboard	decorative wall	tan baseboard	black bird	black handles	wooden holder
10	gold bed	purple clothes	blue doors	wooden pole	wooden cabinet door	wooden tires	black handles	purple shade	red roof

Table 10: Column names indicate either the most frequently occurring objects in images (the top section of the table) or the least frequently occurring ones (the bottom section of the table). Similarities are calculated on the **scene vector space** and objects are ordered from the most similar (0) to the least similar (9).