A Linguistic Analysis of Visually Grounded Dialogues Based on Spatial Expressions

Takuma Udagawa[1] **Takato Yamazaki**[1] **Akiko Aizawa**[1,2]
The University of Tokyo, Tokyo, Japan[1]
National Institute of Informatics, Tokyo, Japan[2]
{takuma_udagawa,takatoy,aizawa}@nii.ac.jp

Abstract

Recent models achieve promising results in visually grounded dialogues. However, existing datasets often contain undesirable biases and lack sophisticated linguistic analyses, which make it difficult to understand how well current models recognize their precise linguistic structures. To address this problem, we make two design choices: first, we focus on OneCommon Corpus (Udagawa and Aizawa, 2019, 2020), a simple yet challenging common grounding dataset which contains minimal bias by design. Second, we analyze their linguistic structures based on *spatial expressions* and provide comprehensive and reliable annotation for 600 dialogues. We show that our annotation captures important linguistic structures including predicate-argument structure, modification and ellipsis. In our experiments, we assess the model's understanding of these structures through reference resolution. We demonstrate that our annotation can reveal both the strengths and weaknesses of baseline models in essential levels of detail. Overall, we propose a novel framework and resource for investigating fine-grained language understanding in visually grounded dialogues.

1 Introduction

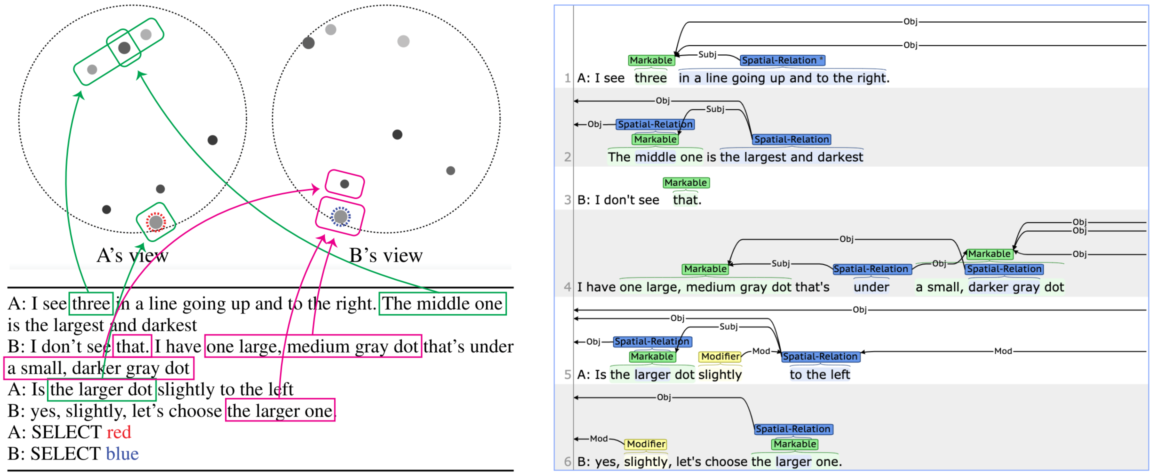
Visual dialogue is the task of holding natural, often goal-oriented conversation in a visual context (Das et al., 2017a; De Vries et al., 2017). This typically involves two types of advanced grounding: *symbol grounding* (Harnad, 1990), which bridges symbolic natural language and continuous visual perception, and *common grounding* (Clark, 1996), which refers to the process of developing mutual understandings through successive dialogues. As noted in Monroe et al. (2017); Udagawa and Aizawa (2019), the *continuous* nature of visual context introduces challenging symbol grounding of nuanced and pragmatic expressions. Some further incorporate *partial observability* where the agents do not share the same context, which introduces complex misunderstandings that need to be resolved through advanced common grounding (Udagawa and Aizawa, 2019; Haber et al., 2019).

Despite the recent progress on these tasks, it remains unclear what types of linguistic structures can (or cannot) be properly recognized by existing models for two reasons. First, existing datasets often contain undesirable biases which make it possible to make correct predictions *without* recognizing the precise linguistic structures (Goyal et al., 2017; Cirik et al., 2018; Agarwal et al., 2020). Second, existing datasets severely lack in terms of sophisticated linguistic analysis, which makes it difficult to understand what types of linguistic structures exist or how they affect model performance.

To address this problem, we make the following design choices in this work:

- We focus on OneCommon Corpus (Udagawa and Aizawa, 2019, 2020), a simple yet challenging collaborative referring task under continuous and partially-observable context. In this dataset, the visual contexts are kept simple and controllable to remove undesirable biases while enhancing linguistic variety. In total, 5,191 natural dialogues are collected and fully annotated with referring expressions (which they called *markables*) and their referents, which can be leveraged for further linguistic analysis.
- To capture the linguistic structures in these dialogues, we propose to annotate *spatial expressions* which play a central role in visually grounded dialogues. We take inspiration from the existing annotation frameworks (Pustejovsky et al., 2011a,b; Petruck and Ellsworth, 2018; Ulinski et al., 2019) but also make several simplifications and modifications to improve coverage,

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 750–765
November 16 - 20, 2020. ©2020 Association for Computational Linguistics



A: I see three in a line going up and to the right. The middle one is the largest and darkest
B: I don't see that. I have one large, medium gray dot that's under a small, darker gray dot
A: Is the larger dot slightly to the left
B: yes, slightly, let's choose the larger one
A: SELECT red
B: SELECT blue

1 A: I see three in a line going up and to the right.
2 The middle one is the largest and darkest
3 B: I don't see that.
4 I have one large, medium gray dot that's under a small, darker gray dot
5 A: Is the larger dot slightly to the left
6 B: yes, slightly, let's choose the larger one.

Figure 1: Example dialogue from OneCommon Corpus with reference resolution annotation (left) and our spatial expression annotation (right). We consider spatial expressions as predicates and annotate their arguments as well as modifiers. For further details of the original dataset and our annotation schema, see Section 3.

efficiency and reliability. [1]

As shown in Figure 1, we consider spatial expressions as *predicates* with existing markables as their *arguments*. We distinguish the argument roles based on *subjects* and *objects* [2] and annotate *modifications* based on nuanced expressions (such as *slightly*). By allowing the arguments to be in previous utterances, our annotation also captures *argument ellipsis* in a natural way.

In our experiments, we focus on reference resolution to study the model's comprehension of these linguistic structures. Since we found the existing baseline to perform relatively poorly, we propose a simple method of incorporating *numerical constraints* in model predictions, which significantly improved its prediction quality.

Based on our annotation, we conduct a series of analyses to investigate whether the model predictions are *consistent* with the spatial expressions. Our main finding is that the model is adept at recognizing entity-level attributes (such as color and size), but mostly fails in capturing inter-entity relations (especially placements): using the terminologies from Landau and Jackendoff (1993), the model can recognize the *what* but not the *where* in spatial language. We also conduct further analyses to investigate the effect of other linguistic factors.

Overall, we propose a novel framework and resource for conducting fine-grained linguistic analyses in visually grounded dialogues. All materials in this work will be publicly available at `https://github.com/Alab-NII/onecommon` to facilitate future model development and analyses.

2 Related Work

Linguistic structure plays a critical role in dialogue research. From theoretical aspects, various dialogue structures have been studied, including discourse structure (Stent, 2000; Asher et al., 2003), speech act (Austin, 1962; Searle, 1969) and common grounding (Clark, 1996; Lascarides and Asher, 2009). In dialogue system engineering, various linguistic structures have been considered and applied, including syntactic dependency (Davidson et al., 2019), predicate-argument structure (PAS) (Yoshino et al., 2011), ellipsis (Quan et al., 2019; Hansen and Søgaard, 2020), intent recognition (Silva et al., 2011; Shi et al., 2016), semantic representation/parsing (Mesnil et al., 2013; Gupta et al., 2018) and frame-based dialogue state tracking (Williams et al., 2016; El Asri et al., 2017). However, most prior work focus on dialogues where information is not grounded in external, perceptual modality such as vision. In this work, we propose an effective method of analyzing linguistic structures in visually grounded dialogues.

Recent years have witnessed an increasing attention in visually grounded dialogues (Zarrieß et al., 2016; de Vries et al., 2018; Alamri et al., 2019; Narayan-Chen et al., 2019). Despite the impressive progress on benchmark scores and model architec-

[1] For instance, we define *spatial expressions* in a broad sense and include spatial attributes (e.g. object size and color) as well as their comparisons.

[2] Our *subject-object* distinction corresponds to other terminologies such as *trajector-landmark* or *figure-ground*.

tures (Das et al., 2017b; Wu et al., 2018; Kottur et al., 2018; Gan et al., 2019; Shukla et al., 2019; Niu et al., 2019; Zheng et al., 2019; Kang et al., 2019; Murahari et al., 2019; Pang and Wang, 2020), there have also been critical problems pointed out in terms of dataset biases (Goyal et al., 2017; Chattopadhyay et al., 2017; Massiceti et al., 2018; Chen et al., 2018; Kottur et al., 2019; Kim et al., 2020; Agarwal et al., 2020) which obscure such contributions. For instance, Cirik et al. (2018) points out that existing dataset of reference resolution may be largely solvable *without* recognizing the full referring expressions (e.g. based on object categories only). To circumvent these issues, we focus on OneCommon Corpus where the visual contents are simple (exploitable categories are removed) and well-balanced (by sampling from uniform distributions) to minimize dataset biases.

Although various probing methods have been proposed for models and datasets in NLP (Belinkov and Glass, 2019; Geva et al., 2019; Kaushik et al., 2020; Gardner et al., 2020; Ribeiro et al., 2020), fine-grained analyses of visually grounded dialogues have been relatively limited. Instead, Kottur et al. (2019) proposed a diagnostic dataset to investigate model's language understanding: however, their dialogues are generated artificially and may not reflect the true nature of visual dialogues. Shekhar et al. (2019) also acknowledges the importance of linguistic analysis but only dealt with coarse-level features that can be computed automatically (such as dialogue topic and diversity). Most similar and related to our research are Yu et al. (2019); Udagawa and Aizawa (2020), where they conducted additional annotation of reference resolution in visual dialogues: however, they still do not capture more sophisticated linguistic structures such as PAS, modification and ellipsis.

Finally, spatial language and cognition have a long history of research (Talmy, 1983; Herskovits, 1987). In computational linguistics, (Kordjamshidi et al., 2010; Pustejovsky et al., 2015) developed the task of spatial role labeling to capture spatial information in text: however, they do not fully address the problem of annotation reliability nor grounding in external visual modality. In computer vision, the VisualGenome dataset (Krishna et al., 2017) provides rich annotation of spatial scene graphs constructed from raw images, but not from raw dialogues. Ramisa et al. (2015); Platonov and Schubert (2018) also worked on modelling spatial prepositions in single sentences. To the best of our knowledge, our work is the first to apply, model and analyze spatial expressions in visually grounded dialogues at full scale.

3 Annotation

3.1 Dataset

Our work extends OneCommon Corpus originally proposed in Udagawa and Aizawa (2019). In this task, two players A and B are given slightly different, overlapping perspectives of a 2-dimensional grid with 7 entities in each view (Figure 1, left). Since only some (4, 5 or 6) of them are in common, this setting is *partially-observable* where complex misunderstandings and partial understandings are introduced. In addition, each entity only has *continuous* attributes (x-value, y-value, color and size), which introduce various nuanced and pragmatic expressions. Note that all entity attributes are generated randomly to enhance linguistic diversity and reduce dataset biases. Under this setting, two players were instructed to converse freely in natural language to coordinate attention on one of the same, common entities. Basic statistics of the dialogues are shown at the top of Table 1 and the frequency of nuanced expressions estimated in Table 2.

Total dialogues	6,760
Avg. utterances per dialogue	4.76
Avg. tokens per utterance	12.37
Successful dialogues	5,191
Annotated markables	40,172
% markables with 1 referent	71.81
% markables with 2 referents	14.85
% markables with $\geq$3 referents	12.03
% markables with 0 referent	1.31

Table 1: OneCommon Corpus statistics.

Nuance Type	% Utterance	Example Usage
Approximation	3.98	**almost** in the middle
Exactness	2.71	**exactly** horizontal
Subtlety	9.37	**slightly** to the right
Extremity	9.35	**very** light dot
Uncertainty	5.79	**Maybe** it's different

Table 2: Estimated frequency of nuanced expressions from Udagawa and Aizawa (2019).

More recently, Udagawa and Aizawa (2020) curated all successful dialogues from the corpus and additionally conducted reference resolution annotation. Specifically, they detected all referring expressions (*markables*) based on minimal noun phrases by trained annotators and identified their referents by multiple crowdworkers (Figure 1 left, highlighted). Both annotations were shown to be reliable with high overall agreement. We show their dataset statistics at the bottom of Table 1.

In this work, we randomly sample 600 dialogues from the latest corpus (5,191 dialogues annotated with reference resolution) to conduct further annotation of spatial expressions.

3.2 Annotation Schema

Our annotation procedure consists of three steps: *spatial expression detection*, *argument identification* and *canonicalization*. Based on these annotation, we conduct fine-grained analyses of the dataset (Subsection 3.3) as well as the baseline models (Subsection 4.2). For further details and examples of our annotation, see Appendix A.

3.2.1 Spatial Expression Detection

Based on the definition from Pustejovsky et al. (2011a,b), spatial expressions are "constructions that make explicit reference to the spatial attributes of an object or spatial relations between objects". [3] We generally follow this definition and detect all spans of spatial attributes and relations in the dialogue. To make the distinction clear, we consider entity-level information like color and size as spatial attributes, and other information such as location and *explicit* attribute comparison as spatial relations. Spatial attributes could be annotated as adjectives ("*dark*"), prepositional phrases ("*of light color*") or noun phrases ("*a black dot*"), while spatial relations could be adjectives ("*lighter*"), prepositions ("*near*"), and so on. We also detect modifiers of spatial expressions based on nuanced expressions (c.f. Table 2).

Although we allow certain flexibility in determining their spans, holistic/dependent expressions (such as "*all shades of gray*", "*sloping up to the right*", "*very slightly*") were instructed to be annotated as a single span. Independent expressions (e.g. connected by conjunctions) could be annotated separately or jointly if they had the same structure (e.g. same arguments and modifiers).

For the sake of efficiency, we do not annotate spatial attributes and their modifiers inside markables (see Figure 1), since their spans and arguments are easy to be detected automatically.

3.2.2 Argument Identification

Secondly, we consider the detected spatial expressions as *predicates* and annotate referring expressions (markables) as their *arguments*. This approach has several advantages: first, it has broad coverage since referring expressions are prevalent in visual dialogues. In addition, by leveraging *exophoric* references which directly bridge natural language and the visual context, we can conduct essential analyses related to symbol grounding across the two modalities (Subsection 4.2).

To be specific, we distinguish the argument roles based on subjects and objects. We allow arguments to be in previous utterances *only if* they are unavailable in the present utterance. Multiple markables can be annotated for the subject/object roles, and no object need to be annotated in cases of spatial attributes, nominal/verbal expressions ("*triangle*", "*clustered*") or *implicit global objects* as in superlatives ("*darkest* (of all)"). If the arguments are indeterminable based on these roles (as in enumeration, e.g. "*From left to right, there are ...*"), they were marked as *unannotatable*. Modificands of the modifiers (which could be either spatial attributes or relations) were also identified in this step.

3.2.3 Canonicalization

Finally, we conduct canonicalization of the spatial expressions and modifiers. Since developing a complete ontology for this domain is infeasible or too expensive, we focus on canonicalizing the central *spatial relations* in this work: we do not canonicalize spatial attributes manually, since we can canonicalize the central spatial attributes automatically (c.f. Subsubsection 4.2.1).

According to Landau (2017), there are 2 classes of relations in spatial language: *functional* class whose core meanings engage force-dynamic relationship (such as *on*, *in*) and *geometric* class whose core meanings engage geometry (such as *left*, *above*). Since functional relations are less common in this dataset and more difficult to define due to their vagueness and context dependence (Platonov and Schubert, 2018), we focus on the following 5 categories of geometric relations and attribute comparisons, including a total of 24 canonical relations which can be defined explicitly.

Direction requires the subjects and objects to be placed in certain orientation: *left*, *right*, *above*, *below*, *horizontal*, *vertical*, *diagonal*.

Proximity is related to distance between subjects, objects or other entities: *near*, *far*, *alone*.

[3] Note that their term *object* corresponds to our term *entity*.

Region restricts the subjects to be in a certain region specified by the objects: *interior*, *exterior*.

Color comparison is related to comparison of color between subjects and objects: *lighter*, *lightest*, *darker*, *darkest*, *same color*, *different color*.

Size comparison is related to comparison of size between subjects and objects: *smaller*, *smallest*, *larger*, *largest*, *same size*, *different size*.

To be specific, we annotate whether each detected spatial relation *implies* any of the 24 canonical relations. Each spatial relation can imply multiple canonical relations (e.g. "on the upper right" implies *right* and *above*) or none (e.g. "triangle" does not imply any of the above relations).

In addition, we define 6 modification types (*subtlety*, *extremity*, *uncertainty*, *certainty*, *neutrality* and *negation*) and canonicalize each modifier into one type. For example, "very slightly" is considered to have the overall type of *subtlety*.

3.3 Results

3.3.1 Annotation Reliability

Annotation	% Agreement	Cohen's κ
Attribute Span	98.5	0.88
Relation Span	95.1	0.87
Modifier Span	99.2	0.86
Subject Ident.	98.8	0.96
Object Ident.	95.9	0.79
Modificand Ident.	99.6	0.98
Relation Canon.	99.7	0.96
Modifier Canon.	87.5	0.83

Table 3: Results of our reliability analysis.

To test the reliability of our annotation, two trained annotators (the authors) independently detected the spatial expressions and modifiers in 50 dialogues. Then, using the 50 dialogues from one of the annotators, two annotators independently conducted argument identification and canonicalization. We show the observed agreement and Cohen's κ (Cohen, 1968) in Table 3.

For span detection, we computed the token level agreement of spatial expressions and modifiers. Despite having certain freedom for determining their spans, we observed very high agreement (including their starting positions, see Appendix B).

For argument identification, we computed the exact match rate of the arguments and modificands. As a result, we observed near perfect agreement for subject/modificand identification. For object identification, the result seems relatively worse: however, upon further inspection, we verified that 73.5% of the disagreements were essentially based on the same markables (e.g. coreferences).

Finally, we observed reasonably high agreement for relation/modifier canonicalization as well. Overall, we conclude that all steps of our annotation can be conducted with high reliability.

3.3.2 Annotation Statistics

	Attribute	Relation
Total	378	4,300
Unique	121	1,139
Avg. per dialogue	0.63	7.17
% inter-utterance subject	1.59	1.37
% inter-utterance object	-	14.65
% no object	-	30.84
% modified	36.51	16.86
% unannotatable	0.79	0.79

Table 4: Statistics of our spatial expression annotation in 600 randomly sampled dialogues.

The basic statistics of our annotation are summarized in Table 4. Note that there are relatively few spatial attributes annotated, since most of them appeared inside the markables (hence not detected manually). However, a large number of spatial relations with non-obvious structures were identified.

In both expressions, we found over 1% of the subjects and 14% of the objects to be present only in previous utterances, which indicates that argument level ellipses are common and need to be resolved in visual dialogues. For spatial relations, about 30% did not have any explicit objects.

Our annotation also verified that a large portion of the spatial expressions (37% for spatial attributes and 17% for relations) accompanied modifiers.

Finally, less than 1% of spatial expressions were *unannotatable* based on our schema, which verifies its broad coverage. Overall, our annotation can capture important linguistic structures of visually grounded dialogues, and it is straightforward to conduct even further analyses (e.g. by focusing on specific canonical relations or modifications).

4 Experiments

4.1 Reference Resolution

Reference resolution is an important subtask of visual dialogue that can be used for probing model's understanding of intermediate dialogue process (Udagawa and Aizawa, 2020). As illustrated in Figure 1 (left), this is a simple task of predicting the referents for each markable based on the *speaker*'s perspective. To collect model predictions for all dialogues, we split the whole dataset into 10 equal-sized bins and use each bin as the test set in 10 rounds of the experiments. For a more detailed setup of our experiments, see Appendix C.

4.1.1 Models

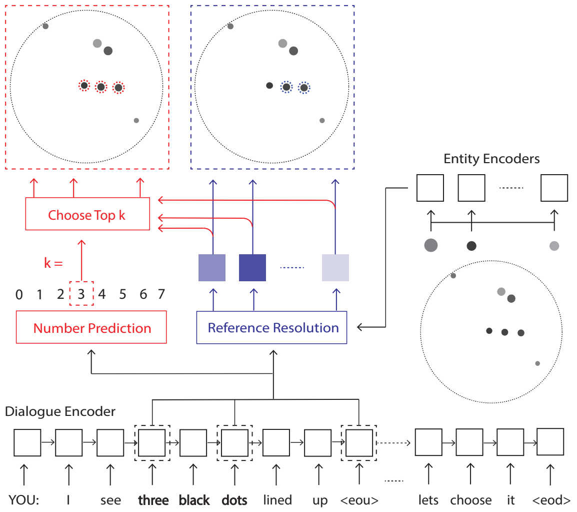


Figure 2: Our model architecture. REF prediction flow is shown in blue and NUMREF prediction flow in red.

As a baseline, we use the REF model proposed in Udagawa and Aizawa (2020). As shown in Figure 2, this model has two encoders: *dialogue encoder* based on a simple GRU (Cho et al., 2014) and *entity encoder* which outputs entity-level representation of the observation based on MLP and relational network (Santoro et al., 2017). To predict the referents, REF takes the GRU's start position of the markable, end position of the markable and end position of the utterance to compute entity-level scores and judge whether each entity is a referent based on logistic regression.

However, since the predictions are made independently for each entity, this model often predicts the wrong number of referents, leading to low performance in terms of exact match rate. To address this issue, we trained a separate module to track the *number* of referents in each markable. We formulate this as a simple classification task between 0, 1, ..., 7, which can be predicted reliably with an average accuracy of 92%. Based on this module's prediction k, we simply take the top k entities with the highest scores as the referents. We refer to this numerically constrained model as NUMREF.

Furthermore, we conduct feature level ablations to study the importance of each feature: for instance, we remove the xy-values from the structured input to ablate the *location* feature.

4.1.2 Results

	Entity-Level Accuracy	Markable-Level Exact Match
REF	85.71±0.23	33.15±1.00
REF−location	84.28±0.27	30.53±0.84
REF−color	83.08±0.32	17.09±1.04
REF−size	83.50±0.22	19.41±0.98
NUMREF	**86.03±0.33**	**54.94±0.76**
NUMREF−location	83.35±0.26	49.77±0.64
NUMREF−color	81.19±0.41	39.74±1.31
NUMREF−size	82.39±0.20	43.40±0.67
Human	96.26	86.90

Table 5: Reference resolution results.

We report the mean and standard deviation of the entity-level accuracy and markable-level exact match rate in Table 5. Compared to REF, our NUMREF model slightly improves the entity-level accuracy and significantly outperforms it in terms of exact match rate, which validates our motivation. From the ablation studies, we can see that all features contribute to the overall performance, but color and size seem to have the largest impact.

However, it is difficult to see how and where these models struggle based on mere accuracy. For further investigation, we need more sophisticated *behavioral testing* (namely black-box testing) to verify whether each model has the capability of recognizing certain concepts or linguistic structures (Ribeiro et al., 2020).

4.2 Model Analysis

To study the current model's strengths and weaknesses in detail, we investigate whether their predictions are *consistent* with the central spatial expressions.

4.2.1 Spatial Attributes

First, we analyze whether the model predictions are consistent with the entity-level spatial attributes. Since most of them were confirmed to appear inside the markables (Subsection 3.3), we automatically detect all expressions of *color* in the markables, plot the distributions of the actual referent color, and compare the results between gold human annotation and model predictions (Figure 3).

From the figure, we can verify that the two distributions look almost identical for the common color expressions, and our NUMREF model seems

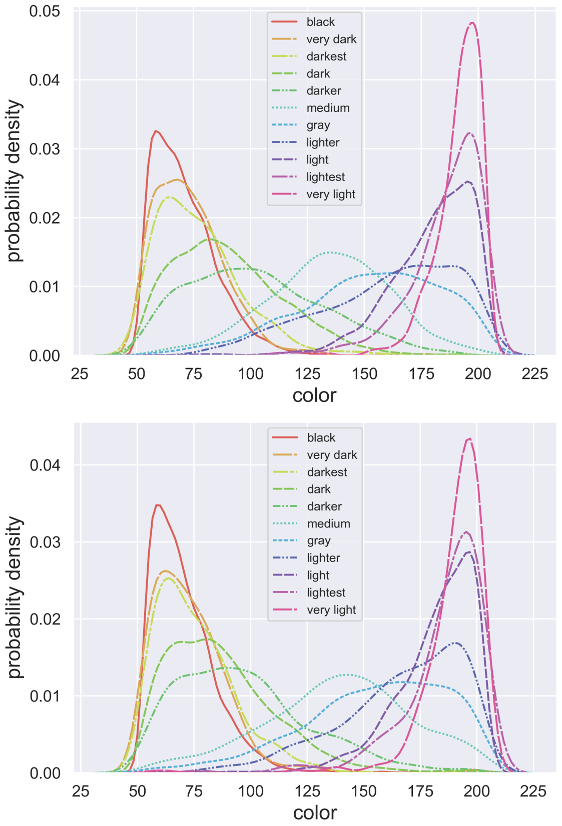


Figure 3: Referent color distributions. Top is human, bottom is `NUMREF` (smaller is darker in color axis).

to capture important characteristics of pragmatic expressions (same expression being used for wide range of colors) and modifications such as neutrality (*medium*) and extremity (*very dark*, *very light*). [4] We observed very similar result with the *size* distributions, which is available in Appendix D.

Based on these results, we argue that the current model can capture entity-level attributes very well, including basic modification.

4.2.2 Spatial Relations

Next, we investigate whether the model predictions are consistent with the central spatial relations. Based on our annotation (Subsection 3.2), we conduct simple tests to check whether the predicted referents satisfy each canonical relation. To be specific, our tests check for two conditions: whether the predictions are *valid* (satisfy the minimal requirements, e.g. at least 2 referents predicted for *near* relation), and if they are valid, whether the predictions actually *satisfy* the canonical relation (e.g. referents are closer than a certain threshold).

Algorithm 1 shows our test for the canonical *left* relation. Note that if no objects are annotated, we simply test whether the referents are on the left side of the player's view. For further details/examples of our canonical relation tests, see Appendix E.

Algorithm 1: Test for *left* relation

```
Input: subject referents $\mathcal{S}$, object referents $\mathcal{O}$,
       boolean $no\_object$
Output: boolean $satisfy$, boolean $valid$
if $no\_object$ then
    $valid \leftarrow |\mathcal{S}| > 0$
    $satisfy \leftarrow valid \wedge mean(\mathcal{S}.x) < 0$
else
    $valid \leftarrow |\mathcal{S}| > 0 \wedge |\mathcal{O}| > 0$
    $satisfy \leftarrow valid \wedge mean(\mathcal{S}.x) < mean(\mathcal{O}.x)$
return $satisfy$, $valid$
```

The results of our tests are summarized in Table 6. We also compare with the feature ablated models to estimate the test cases which can be satisfied *without* using the corresponding features, i.e. location for *direction/proximity/region* categories, color for *color comparison*, and size for *size comparison*.

First, we can verify that human annotation passes most of our tests, which is an important evidence of the *validity* of our annotations and tests. We also confirmed that `REF` models often make *invalid* predictions with overall poor performance, which is consistent with our expectation.

In *direction*, *proximity* and *region* categories, we found that `NUMREF` model performs on par or only marginally better than its ablated version (and even underperforms it for simple relations like *right* and *above*): these results indicate that current model is still incapable of leveraging locational features to make more consistent predictions. [5]

In *color/size comparison*, `NUMREF` performs reasonably well, outperforming all other models: this indicates that the model can not only capture but also *compare* entity-level attributes to a certain extent. However, there is still room left for improvement in almost all relations. It is also worth noting that *size comparison* may be easier, as the range of size is limited (only *6* compared to *150* for color).

Overall, we conclude that current models still struggle in capturing most of the inter-entity relations, especially those related to placements.

4.2.3 Further Analyses

Finally, we conduct further analyses to study other linguistic factors that affect model performance.

[4] Spatial attributes with nuances of subtlety (such as *slightly dark*) were relatively rare and omitted in the figure.

[5] For relations like *far* and *different color*, ablated model may be better simply because referents tend to be more distant/dissimilar when predictions are closer to random.

Models			REF		REF-abl		NUMREF		NUMREF-abl		Human	
Category	Relation	# Cases	satisfy	valid	satisfy	valid	satisfy	valid	satisfy	valid	satisfy	valid
Direction	*left*	412	23.5	32.3	21.1	28.9	**67.0**	**99.5**	62.4	**99.5**	95.9	97.6
	right	468	28.0	35.5	24.6	30.8	67.3	**98.7**	**68.2**	**98.7**	95.3	96.4
	above	514	28.6	37.4	24.7	33.1	65.2	99.2	**66.5**	**99.4**	96.7	98.6
	below	444	25.2	34.5	21.6	27.9	**66.0**	**99.1**	62.2	**99.1**	96.4	96.8
	horizontal	37	54.1	70.3	27.0	59.5	**59.5**	**100.0**	51.4	97.3	91.9	100.0
	vertical	46	37.0	73.9	23.9	54.3	43.5	**95.7**	**45.7**	**95.7**	82.6	100.0
	diagonal	50	48.0	74.0	30.0	50.0	**60.0**	**98.0**	**60.0**	**98.0**	90.0	100.0
	All	1,971	27.8	37.6	23.4	31.9	**65.5**	**99.0**	64.1	**99.0**	95.5	97.6
Proximity	*near*	271	49.4	61.3	29.9	49.1	**77.1**	94.5	56.1	**95.2**	95.2	96.7
	far	27	29.6	40.7	33.3	40.7	77.8	**100.0**	**92.6**	**100.0**	96.3	96.3
	alone	111	36.9	44.1	45.0	54.1	**68.5**	**94.6**	67.6	**94.6**	91.9	94.6
	All	409	44.7	55.3	34.2	49.9	**74.8**	94.9	61.6	**95.4**	94.4	96.1
Region	*interior*	135	38.5	52.6	27.4	39.3	**62.2**	93.3	58.5	**94.1**	96.3	100.0
	exterior	62	40.3	48.4	40.3	53.2	80.6	**98.4**	**87.1**	**98.4**	98.4	98.4
	All	197	39.1	51.3	31.5	43.7	**68.0**	94.9	67.5	**95.4**	97.0	99.5
Color	*lighter*	147	23.1	25.9	6.8	8.2	**84.4**	**100.0**	57.1	99.3	97.3	98.0
	lightest	42	45.2	66.7	14.3	33.3	**61.9**	**100.0**	31.0	**100.0**	83.3	100.0
	darker	171	24.0	26.3	7.0	10.5	**83.0**	**99.4**	53.2	**99.4**	95.9	98.8
	darkest	48	56.2	64.6	14.6	33.3	**66.7**	**100.0**	35.4	**100.0**	89.6	97.9
	same	50	12.0	30.0	8.0	30.0	**40.0**	**88.0**	32.0	86.0	92.0	96.0
	different	14	64.3	71.4	71.4	71.4	64.3	**100.0**	**78.6**	92.9	92.9	100.0
	All	472	28.8	35.4	10.4	18.0	**74.8**	**98.5**	49.2	97.9	94.1	98.3
Size	*smaller*	213	27.7	31.5	7.5	9.9	**80.8**	**100.0**	59.6	**100.0**	98.6	99.5
	smallest	52	71.2	73.1	21.2	34.6	**86.5**	**98.1**	48.1	**98.1**	92.3	98.1
	larger	238	23.1	28.6	9.7	16.0	**73.5**	**99.6**	48.7	**99.6**	98.3	98.3
	largest	61	52.5	60.7	11.5	24.6	**73.8**	**100.0**	39.3	**100.0**	96.7	100.0
	same	103	34.0	42.7	18.4	27.2	**80.6**	88.3	65.0	**91.3**	98.1	100.0
	different	12	75.0	75.0	66.7	66.7	**91.7**	**91.7**	83.3	83.3	91.7	91.7
	All	679	33.4	38.7	12.4	18.9	**78.2**	97.8	54.3	**98.1**	97.6	99.0

Table 6: Canonical relation test results. We compute the *satisfy* and *valid* rate of the predictions for each canonical relation. Best scores of the models are in bold (-abl shows the corresponding feature ablated results).

Linguistic Factors	# Cases	NUMREF	Human
strong modification	149	76.51	95.97
neutral	3,094	70.46	95.77
weak modification	490	66.12	95.10
inter-utterance subject	14	57.14	92.86
inter-utterance object	265	72.08	94.72
no object	1,127	74.45	92.99
ignorable object	1,805	69.64	97.23
unignorable object	796	65.33	96.11
All	3,728	70.17	95.71

Table 7: Satisfy rate classified by linguistic factors.

Table 7 shows the results of our relation tests classified by notable linguistic structures.

In terms of modification, we can confirm that human performance is consistently high, while the model performs best for strong modification (modification types of *extremity* or *certainty*), decently for neutrals (*neutrality* or no modification), and worst on weak modification (*subtlety* or *uncertainty*). This indicates that large, conspicuous features are easier for the model to capture compared to small or more ambiguous features.

In terms of subject/object properties, human performance is also consistently high. In contrast, model performance is significantly worse for subject ellipsis (*inter-utterance subject*), while remaining high for object ellipsis and *no object* cases.

We also hypothesize that a large portion of the relations can actually be satisfied *without* considering the objects, e.g. by simply predicting very dark dots as the subjects when the relation is *darker* or *darkest*. To distinguish such easy cases, we consider a relation as *ignorable object* if the relation can be satisfied even if we ignore the objects (i.e. remove all object relations) based on gold referents. Our result verifies that there are indeed many cases of *ignorable object*, and they seem slightly easier for the model to satisfy.

Models		NUMREF		Human	
value	mod-type	diff.	# valid	diff.	# valid
xy-value	strong	86.06	39	89.15	37
	neutral	80.92	1,586	73.52	1,558
	weak	80.35	200	53.53	198
color	strong	66.23	15	91.80	15
	neutral	56.98	234	60.14	232
	weak	37.73	68	28.55	66
size	strong	3.60	8	4.29	8
	neutral	2.67	337	2.70	320
	weak	1.95	105	1.58	104

Table 8: Absolute difference in comparative relations (number of valid predictions shown in shade).

In Table 8, we study the effect of modification based on the *absolute difference* between subject and object features in comparative relations. [6]

In human annotation, the absolute difference naturally increases as the modification gets stronger. While model predictions also show this tendency, their results seem less sensitive to modification (particularly for locational features, i.e. xy-value) and may not be reflecting their full effect.

5 Discussion and Conclusion

In this work, we focused on the recently proposed OneCommon Corpus as a suitable testbed for fine-grained language understanding in visually grounded dialogues. To analyze its linguistic structures, we proposed a novel framework of annotating spatial expressions in visual dialogues. We showed that our annotation can be conducted reliably and efficiently by leveraging referring expressions prevalent in visual dialogues, while capturing important linguistic structures such as PAS, modification and ellipsis. Although our current analysis is limited to this domain, we expect that upon appropriate definition of spatial expressions, argument roles and canonicalization, the general approach can be applied to a wider variety of domains: adapting and validating our approach in different domains (especially with more realistic visual contexts) are left as future work.

Secondly, we proposed a simple idea of incorporating *numerical constraints* to improve exophoric reference resolution. We expect that a similar approach of identifying and incorporating semantic constraints (e.g. coreferences and spatial constraints) is a promising direction to improve the model's performance even further.

Finally, we demonstrated the advantages of our annotation for investigating the model's understanding of visually grounded dialogues. Our tests are completely agnostic to the models and only require referent predictions made by each model. By designing simple tests like ours (Subsubsection 4.2.1/4.2.2), we can diagnose the model's performance at the granularity of canonical attributes/relations under consideration: such analyses are easy to extend (by adding more tests) and critical for verifying what capabilities current models have (or do not have). Based on further analyses (Subsubsection 4.2.3), we also revealed various linguistic structures that affect model performance: we expect that capturing and studying such effects will be essential for advanced model probing in visual dialogue research.

Overall, we expect our framework and resource to be fundamental for conducting sophisticated linguistic analyses of visually grounded dialogues.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 18H03297 and NEDO SIP-2 "Big-data and AI-enabled Cyberspace Technologies." We also thank the anonymous reviewers for their valuable suggestions and comments.

References

Shubham Agarwal, Trung Bui, Joon-Young Lee, Ioannis Konstas, and Verena Rieser. 2020. History for visual dialog: Do we really need it? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8182–8197, Online. Association for Computational Linguistics.

Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Anoop Cherian, Irfan Essa, Dhruv Batra, Tim K Marks, Chiori Hori, Peter Anderson, et al. 2019. Audio visual scene-aware dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7558–7567.

Nicholas Asher, Nicholas Michael Asher, and Alex Lascarides. 2003. *Logics of conversation*. Cambridge University Press.

[6] *Left/right* for x-value, *above/below* for y-value, *lighter/darker* for color and *smaller/larger* for size.

John Langshaw Austin. 1962. *How to do things with words*. Oxford university press.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics (TACL)*, 7:49–72.

Prithvijit Chattopadhyay, Deshraj Yadav, Viraj Prabhu, Arjun Chandrasekaran, Abhishek Das, Stefan Lee, Dhruv Batra, and Devi Parikh. 2017. Evaluating visual conversational agents via cooperative human-ai games. In *Proceedings of the Fifth AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*.

Hongge Chen, Huan Zhang, Pin-Yu Chen, Jinfeng Yi, and Cho-Jui Hsieh. 2018. Attacking visual language grounding with adversarial examples: A case study on neural image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2587–2597, Melbourne, Australia. Association for Computational Linguistics.

Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111. Association for Computational Linguistics.

Volkan Cirik, Louis-Philippe Morency, and Taylor Berg-Kirkpatrick. 2018. Visual referring expression recognition: What do systems actually learn? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 781–787, New Orleans, Louisiana. Association for Computational Linguistics.

Herbert H Clark. 1996. *Using language*. Cambridge university press.

Jacob Cohen. 1968. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017a. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2.

Abhishek Das, Satwik Kottur, Jose M. F. Moura, Stefan Lee, and Dhruv Batra. 2017b. Learning cooperative visual dialog agents with deep reinforcement learning. In *The IEEE International Conference on Computer Vision (ICCV)*.

Sam Davidson, Dian Yu, and Zhou Yu. 2019. Dependency parsing for spoken dialog systems. 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 1513–1519, Hong Kong, China. Association for Computational Linguistics.

Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 2017. Guesswhat?! visual object discovery through multi-modal dialogue. In *Proc. of CVPR*.

Layla El Asri, Hannes Schulz, Shikhar Sharma, Jeremie Zumer, Justin D. Harris, Emery Fine, Rahul Mehrotra, and Kaheer Suleman. 2017. Frames: A corpus for adding memory to goal-oriented dialogue systems. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 207–219. Association for Computational Linguistics.

Zhe Gan, Yu Cheng, Ahmed Kholy, Linjie Li, Jingjing Liu, and Jianfeng Gao. 2019. Multi-step reasoning via recurrent dual attention for visual dialog. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6463–6474, Florence, Italy. Association for Computational Linguistics.

Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, et al. 2020. Evaluating nlp models via contrast sets. *arXiv preprint arXiv:2004.02709*.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. 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 1161–1166, Hong Kong, China. Association for Computational Linguistics.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913.

Sonal Gupta, Rushin Shah, Mrinal Mohit, Anuj Kumar, and Mike Lewis. 2018. Semantic parsing for task oriented dialog using hierarchical representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2787–2792, Brussels, Belgium. Association for Computational Linguistics.

Janosch Haber, Tim Baumgärtner, Ece Takmaz, Lieke Gelderloos, Elia Bruni, and Raquel Fernández. 2019. The PhotoBook dataset: Building common ground through visually-grounded dialogue. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1895–1910, Florence, Italy. Association for Computational Linguistics.

Victor Petrén Bach Hansen and Anders Søgaard. 2020. What do you mean 'why?': Resolving sluices in conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7887–7894.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346.

Annette Herskovits. 1987. *Language and spatial cognition*. Cambridge university press.

Gi-Cheon Kang, Jaeseo Lim, and Byoung-Tak Zhang. 2019. Dual attention networks for visual reference resolution in visual dialog. 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 2024–2033, Hong Kong, China. Association for Computational Linguistics.

Divyansh Kaushik, Eduard Hovy, and Zachary C Lipton. 2020. Learning the difference that makes a difference with counterfactually-augmented data. *International Conference on Learning Representations (ICLR)*.

Hyounghun Kim, Hao Tan, and Mohit Bansal. 2020. Modality-balanced models for visual dialogue. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8091–8098.

Parisa Kordjamshidi, Martijn Van Otterlo, and Marie-Francine Moens. 2010. Spatial role labeling: Task definition and annotation scheme. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Satwik Kottur, Jose M. F. Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. 2018. Visual coreference resolution in visual dialog using neural module networks. In *The European Conference on Computer Vision (ECCV)*.

Satwik Kottur, José M. F. Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. 2019. CLEVR-dialog: A diagnostic dataset for multi-round reasoning in visual dialog. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 582–595, Minneapolis, Minnesota. 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, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73.

Barbara Landau. 2017. Update on "what" and "where" in spatial language: A new division of labor for spatial terms. *Cognitive science*, 41:321–350.

Barbara Landau and Ray Jackendoff. 1993. "what" and "where" in spatial language and spatial cognition. *Behavioral and brain sciences*, 16(2):217–238.

Alex Lascarides and Nicholas Asher. 2009. Agreement, disputes and commitments in dialogue. *Journal of semantics*, 26(2):109–158.

Daniela Massiceti, Puneet K Dokania, N Siddharth, and Philip HS Torr. 2018. Visual dialogue without vision or dialogue. *arXiv preprint arXiv:1812.06417*.

Grégoire Mesnil, Xiaodong He, Li Deng, and Yoshua Bengio. 2013. Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding. In *Interspeech*, pages 3771–3775.

Will Monroe, Robert X. D. Hawkins, Noah D. Goodman, and Christopher Potts. 2017. Colors in context: A pragmatic neural model for grounded language understanding. *Transactions of the Association for Computational Linguistics*, 5:325–338.

Vishvak Murahari, Dhruv Batra, Devi Parikh, and Abhishek Das. 2019. Large-scale pretraining for visual dialog: A simple state-of-the-art baseline. *arXiv preprint arXiv:1912.02379*.

Anjali Narayan-Chen, Prashant Jayannavar, and Julia Hockenmaier. 2019. Collaborative dialogue in Minecraft. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5405–5415, Florence, Italy. Association for Computational Linguistics.

Yulei Niu, Hanwang Zhang, Manli Zhang, Jianhong Zhang, Zhiwu Lu, and Ji-Rong Wen. 2019. Recursive visual attention in visual dialog. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wei Pang and Xiaojie Wang. 2020. Visual dialogue state tracking for question generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11831–11838.

Miriam R. L. Petruck and Michael J. Ellsworth. 2018. Representing spatial relations in FrameNet. In *Proceedings of the First International Workshop on Spatial Language Understanding*, pages 41–45, New Orleans. Association for Computational Linguistics.

Georgiy Platonov and Lenhart Schubert. 2018. Computational models for spatial prepositions. In *Proceedings of the First International Workshop on Spatial Language Understanding*, pages 21–30.

James Pustejovsky, Parisa Kordjamshidi, Marie-Francine Moens, Aaron Levine, Seth Dworman, and Zachary Yocum. 2015. Semeval-2015 task 8:

Spaceeval. In *Proceedings of the 9th International Workshop on Semantic Evaluation (semeval 2015)*, pages 884–894. ACL.

James Pustejovsky, Jessica L Moszkowicz, and Marc Verhagen. 2011a. Iso-space: The annotation of spatial information in language. In *Proceedings of the Sixth Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation*, volume 6, pages 1–9.

James Pustejovsky, Jessica L Moszkowicz, and Marc Verhagen. 2011b. Using iso-space for annotating spatial information. In *Proceedings of the International Conference on Spatial Information Theory*.

Jun Quan, Deyi Xiong, Bonnie Webber, and Changjian Hu. 2019. GECOR: An end-to-end generative ellipsis and co-reference resolution model for task-oriented dialogue. 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 4547–4557, Hong Kong, China. Association for Computational Linguistics.

Arnau Ramisa, Josiah Wang, Ying Lu, Emmanuel Dellandrea, Francesc Moreno-Noguer, and Robert Gaizauskas. 2015. Combining geometric, textual and visual features for predicting prepositions in image descriptions. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 214–220, Lisbon, Portugal. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Tim Lillicrap. 2017. A simple neural network module for relational reasoning. In *Advances in neural information processing systems*, pages 4967–4976.

John R Searle. 1969. *Speech acts: An essay in the philosophy of language*, volume 626. Cambridge university press.

Ravi Shekhar, Aashish Venkatesh, Tim Baumgärtner, Elia Bruni, Barbara Plank, Raffaella Bernardi, and Raquel Fernández. 2019. Beyond task success: A closer look at jointly learning to see, ask, and GuessWhat. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2578–2587, Minneapolis, Minnesota. Association for Computational Linguistics.

Yangyang Shi, Kaisheng Yao, Le Tian, and Daxin Jiang. 2016. Deep LSTM based feature mapping for query classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1501–1511, San Diego, California. Association for Computational Linguistics.

Pushkar Shukla, Carlos Elmadjian, Richika Sharan, Vivek Kulkarni, Matthew Turk, and William Yang Wang. 2019. What should I ask? using conversationally informative rewards for goal-oriented visual dialog. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6442–6451, Florence, Italy. Association for Computational Linguistics.

Joao Silva, Luísa Coheur, Ana Cristina Mendes, and Andreas Wichert. 2011. From symbolic to subsymbolic information in question classification. *Artificial Intelligence Review*, 35(2):137–154.

Amanda Stent. 2000. Rhetorical structure in dialog. In *INLG'2000 Proceedings of the First International Conference on Natural Language Generation*, pages 247–252, Mitzpe Ramon, Israel. Association for Computational Linguistics.

Leonard Talmy. 1983. How language structures space. In *Spatial orientation*, pages 225–282. Springer.

Takuma Udagawa and Akiko Aizawa. 2019. A natural language corpus of common grounding under continuous and partially-observable context. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7120–7127.

Takuma Udagawa and Akiko Aizawa. 2020. An annotated corpus of reference resolution for interpreting common grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9081–9089.

Morgan Ulinski, Bob Coyne, and Julia Hirschberg. 2019. SpatialNet: A declarative resource for spatial relations. In *Proceedings of the Combined Workshop on Spatial Language Understanding (SpLU) and Grounded Communication for Robotics (RoboNLP)*, pages 61–70, Minneapolis, Minnesota. Association for Computational Linguistics.

Harm de Vries, Kurt Shuster, Dhruv Batra, Devi Parikh, Jason Weston, and Douwe Kiela. 2018. Talk the walk: Navigating new york city through grounded dialogue. *arXiv preprint arXiv:1807.03367*.

Jason Williams, Antoine Raux, and Matthew Henderson. 2016. The dialog state tracking challenge series: A review. *Dialogue & Discourse*, 7(3):4–33.

Qi Wu, Peng Wang, Chunhua Shen, Ian Reid, and Anton van den Hengel. 2018. Are you talking to me? reasoned visual dialog generation through adversarial learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

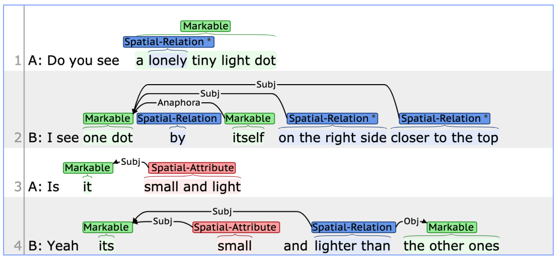
Koichiro Yoshino, Shinsuke Mori, and Tatsuya Kawahara. 2011. Spoken dialogue system based on information extraction using similarity of predicate argument structures. In *Proceedings of the SIGDIAL 2011 Conference*, pages 59–66. Association for Computational Linguistics.

Xintong Yu, Hongming Zhang, Yangqiu Song, Yan Song, and Changshui Zhang. 2019. What you see is what you get: Visual pronoun coreference resolution in dialogues. 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 5123–5132, Hong Kong, China. Association for Computational Linguistics.

Sina Zarrieß, Julian Hough, Casey Kennington, Ramesh Manuvinakurike, David DeVault, Raquel Fernández, and David Schlangen. 2016. PentoRef: A corpus of spoken references in task-oriented dialogues. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 125–131, Portorož, Slovenia. European Language Resources Association (ELRA).

Zilong Zheng, Wenguan Wang, Siyuan Qi, and Song-Chun Zhu. 2019. Reasoning visual dialogs with structural and partial observations. In *Computer Vision and Pattern Recognition (CVPR), 2019 IEEE Conference on*.

A Annotation Examples and Details



A: Do you see a lonely tiny light dot
B: I see one dot by itself on the right side closer to the top
A: Is it small and light
B: Yeah its small and lighter than the other ones

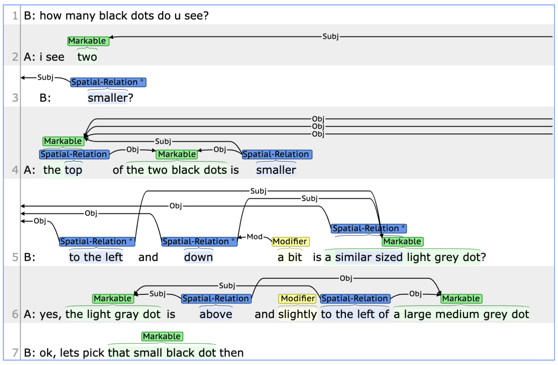
Figure 4: Example with spatial attributes.

Here, we show additional examples of our spatial expression annotation. In Figure 4, we show an example dialogue annotated with *spatial attributes* (colored in red). Since our goal is *not* to achieve strict inter-annotator agreement but to conduct efficient and useful analysis, we allow certain flexibility in determining the spans of spatial expressions: for instance, the coordinated spatial expression ("small and light") can be annotated as a single expression or as different expressions ("small and light"). Copulas (*is*, *being*), articles (*a*, *the*), particles (*to*, *with*) and modifiers were allowed to be either omitted or included in spatial expressions. Spans were allowed to be non-contiguous, but must be annotated at the token level and restricted to be within a single utterance. Note that spatial attributes (*tiny*, *light*) in the first markable ("a lonely tiny light dot") are not annotated, since they are inside the markable and their spans and subjects are relatively obvious.

In terms of argument identification, we prioritize markables in the following manner:

1. Markables in the present utterance (i.e. same utterance as the spatial expression).
2. Markables in the closest previous utterance of the *same speaker*.
3. Markables in the closest previous utterance of *different speakers*.

As long as these priorities are satisfied, we did not distinguish between coreferences. Furthermore, for object identification, we did not distinguish between markables which include/exclude subject referents: for example, the object markable for *lighter* in "I have [three dots], [two] dark and [one] lighter" could be either *three dots* or *two*.

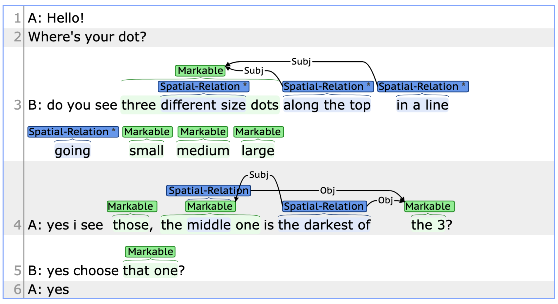


B: how many black dots do u see?
A: i see two
B: smaller?
A: the top of the two black dots is smaller
B: to the left and down a bit is a similar sized light grey dot?
A: yes, the light gray dot is above and slightly to the left of a large medium grey dot
B: ok, lets pick that small black dot then

Figure 5: Example with subject ellipsis.

In Figure 5, we show an example dialogue where the subject markable only appears in the previous utterance ("smaller?" in B's utterance), which demonstrates the case of *subject ellipsis*. Note that since we only detect expressions that contain *specific spatial information* of the visual context, we do not annotate *black dots* in the first interrogative utterance ("how many black dots do u see?").

In Figure 6, we show an example dialogue with *unannotatable* relation ("going [small], [medium], [large]") which cannot be captured based on the simple argument roles of subjects and objects. In general, similar strategies of enumeration are difficult to be captured, as well as predications with *exceptions* (such as "[All dots] are dark except [one



A: Hello!
Where's your dot?
B: do you see three different size dots along the top in a line going small medium large
A: yes i see those, the middle one is the darkest of the 3?
B: yes choose that one?
A: yes

Figure 6: Example with unannotatable relation.

dot]") or cases with *bundled* subjects ("[Two dots] are dark and darker").

Finally, we only annotate *explicit* spatial attributes and relations: therefore, we do not annotate *implicit* relations such as *darker* in "One is dark and the other is light gray", although it is inferable. When the spans are difficult to annotate, annotators were encouraged to make the best effort to capture the constructions which refer to specific spatial information.

B Annotation Results

Annotation	% Agreement	Cohen's κ
Attribute Start	98.5	0.84
Relation Start	95.1	0.77
Modifier Start	98.7	0.82

Table 9: Additional results of our reliability analysis.

In Table 9, we show the results of token level agreement for the *starting positions* of spatial expressions and modifiers. Despite having certain freedom as discussed in Appendix A, we can verify that these also have reasonably high agreement.

	Attribute	Relation
% mod-subtlety	1.06	8.12
% mod-extremity	9.00	2.16
% mod-uncertainty	7.41	4.26
% mod-certainty	0.27	1.40
% mod-neutrality	19.31	0.67
% mod-negation	0.53	0.42

Table 10: Additional statistics of our spatial expression annotation.

In Table 10, we show the frequency of each modification types. Based on these results, we can see that *neutrality* is the most common type of modification for spatial attributes (as in *medium gray*, *medium sized*), and *subtlety* and *uncertainty* to be the most common types for spatial relations. It is interesting to note that the frequencies of modification types vary significantly with spatial attributes and relations, except for *negation*.

In Table 11 and 12, we show the statistics and examples of canonical relations and modification types annotated for our analyses. Note that a single expression can imply multiple canonical relations (e.g. "identical looking" implies *same color* and *same size*) or no canonical relation at all (e.g. "forms a triangle"). In contrast, a modifier can have only one modification type: for instance, *almost exactly* is considered to have the overall modification type of *certainty*.

C Experiment Setup

We use the dataset, baselines, hyperparameters and evaluation metrics publicly available at https://github.com/Alab-NII/onecommon.

In order to collect model predictions for all dialogues and markables, we randomly split the whole dataset into 10 equal sized bins z_i ($i \in \{0, 1, 2, ..., 9\}$) and at each round $r \in \{0, 1, 2, ..., 9\}$ we use $z_{r \pmod{10}}$, $z_{r+1 \pmod{10}}$, ..., $z_{r+7 \pmod{10}}$ for model training, $z_{r+8 \pmod{10}}$ for validation, and $z_{r+9 \pmod{10}}$ for testing. We report the mean and standard deviation of the entity-level accuracy and markable-level exact match rate in these 10 rounds of the experiments.

In our NUMREF model, we train a separate module for predicting the number of referents based on a simple MLP (single layer, 256 hidden units). Reference resolution and number prediction are trained jointly with the loss weighted by 32:1. We conducted minimal hyperparameter tuning since the results did not change dramatically.

D Size Distribution Plots

Figure 7 shows the referent *size* distributions based on human annotation (top) and NUMREF predictions (bottom). We can verify that the two distributions look almost identical for all common expressions, as observed in the color distributions.

E Canonical Relation Tests

For canonical relation tests, we only use relations that are *not negated* and have all arguments in the *same speaker*'s utterances (so that referent predictions are based on the same player's observation). As illustrative examples, we show the algorithms for testing the *horizontal* relation (Algorithm 2),

Category	Relation	Unique	Examples
Direction	*left*	150	to the left (78), on the left (35), left most (5), furthest left (2)
	right	192	to the right (120), on the right (38), lower right (6), to the northeast (1)
	above	190	above (118), top (92), on top (33), up (17), higher (10), just above (4)
	below	179	below (88), bottom (56), lower (38), down (14), lowest (7), beneath (4)
	horizontal	19	horizontal (12), in a horizontal line (4), side by side (3), across from (1)
	vertical	29	vertical (7), on top of (5), on a vertical line (4), aligned vertically with (1)
	diagonal	38	diagonal (5), in a diagonal line (5), sloping down to the right (1), slanted (1)
Proximity	*near*	59	close together (63), cluster (32), next to (28), close to (22), near (13)
	far	21	far (5), away from (4), set apart from (1), a ways above (1), a distance from (1)
	alone	13	by (38), lonely (30), alone (21), lonesome (1), isolated (1)
Region	*interior*	47	middle (41), in the middle (19), between (9), in the center of (2)
	exterior	46	close to the border (5), all around (1), on the outside of (1), surrounding (1)
Color	*lighter*	22	lighter (102), lighter than (10), lighter gray (8), larger lighter (4)
	lightest	11	lightest (28), lightest shade (3), the lightest of (2), lightest and smallest (2)
	darker	30	darker (130), darker than (16), smaller and darker (4), darker in color (3)
	darkest	10	darkest (40), smallest darkest (2), the darkest of (1), darkest/largest of (1)
	same	9	same color (9), identical looking (2), similar shades (1), equally black (1)
	different	11	different shades (3), different sizes and shades (2), of varying shades (1)
Size	*smaller*	17	smaller (209), smaller than (5), smaller and lighter (4), tinier (1)
	smallest	8	smallest (40), tiniest (4), smallest darkest (2), smallest of (1)
	larger	32	larger (178), bigger than (7), larger in size (2), double the size of (1)
	largest	10	largest (41), biggest (11), largest of (2), biggest one of (1)
	same	32	same size (24), same sized (12), similar in size (5), identical in size (3)
	different	8	different sizes (3), of different sizes (1), varying sizes (1), opposite in sizes (1)

Table 11: Unique numbers and examples of spatial relations which imply each canonical relation (frequencies shown in parentheses).

Modification	Unique	Examples
Subtlety	27	slightly (235), a little (48), a bit (35), a tiny bit (8), very slightly (5)
Extremity	15	very (87), much (17), pretty (8), quite (3), really (2)
Uncertainty	36	almost (85), about (49), kind of (23), smallish (6), not completely (3)
Certainty	13	directly (28), exactly (2), perfect (2), almost exactly (2)
Neutrality	16	medium (59), med (9), fairly (4), mid-size (3), slightly medium (1)
Negation	4	not (17), isn't (1), not perceptibly (1)

Table 12: Unique numbers and examples of modifiers with each modification type (frequencies in parentheses).

near relation (Algorithm 3), *interior* relation (Algorithm 4) and *same color* relation (Algorithm 5). Note that each algorithm can take a variety of inputs, such as *all referents* including both subjects and objects ($\mathcal{A}$) or *all observable entities* of the player ($\mathcal{E}$).

Algorithm 2: Test for *horizontal* relation

```
Input: all referents $\mathcal{A}$
Output: boolean $satisfy$, boolean $valid$
$valid \leftarrow |\mathcal{A}| > 1$
if $valid$ then
    // Conduct linear regression and
       check if coeficient is small
    $reg.fit(\mathcal{A})$
    $satisfy \leftarrow reg.coef < \frac{1}{3}$
else
    $satisfy \leftarrow$ False
return $satisfy$, $valid$
```

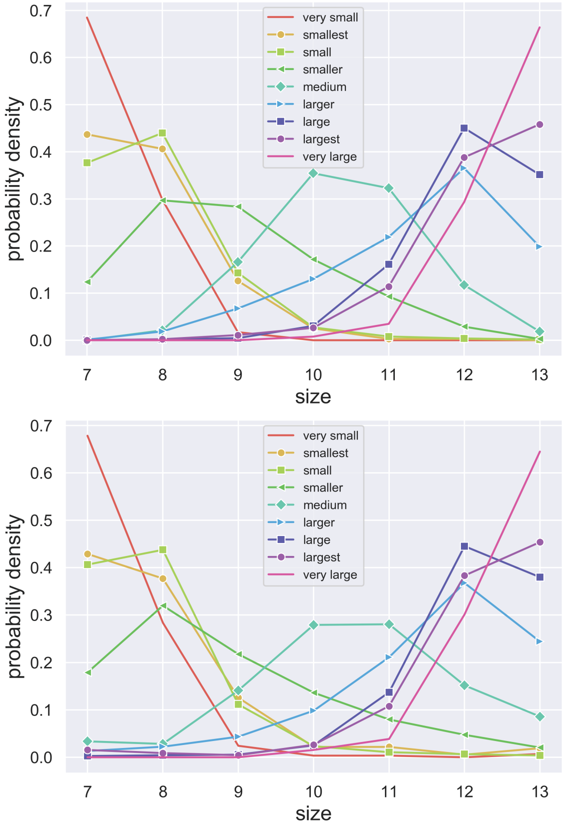


Figure 7: Referent size distributions (top is human, bottom is NUMREF).

```
Algorithm 3: Test for near relation
Input: all referents $\mathcal{A}$, observable entities $\mathcal{E}$
Output: boolean $satisfy$, boolean $valid$
$valid \leftarrow |\mathcal{A}| > 1$
if $valid$ then
    // Compute distance for every
       pair in the set
    $A\_dists \leftarrow dist(x, y)$ for $x, y$ in
     $combination(\mathcal{A})$
    $E\_dists \leftarrow dist(x, y)$ for $x, y$ in
     $combination(\mathcal{E})$
    // Check if mean distance is
       smaller
    $satisfy \leftarrow valid \land$
     $mean(A\_dists) < mean(E\_dists)$
else
    $satisfy \leftarrow$ False
return $satisfy, valid$
```

```
Algorithm 4: Test for interior relation
Input: subject referents $\mathcal{S}$, object referents $\mathcal{O}$,
       boolean $no\_object$
Output: boolean $satisfy$, boolean $valid$
if $no\_object$ then
    // If any subject referent is
       far from the center, satisfy
       is False
    $valid \leftarrow |\mathcal{S}| > 0$
    $satisfy \leftarrow valid$
    $center \leftarrow (0, 0)$
    for $s \in \mathcal{S}$ do
        if $dist(s, center) > 120$ then
            $satisfy \leftarrow$ False
else
    // If any subject referent is
       outside the region of
       objects, satisfy is False
    $valid \leftarrow |\mathcal{S}| > 0 \land |\mathcal{O}| > 1$
    $satisfy \leftarrow valid$
    for $s \in \mathcal{S}$ do
        if $(s.x < min(\mathcal{O}.x) \lor max(\mathcal{O}.x) < s.x) \land$
        $(s.y < min(\mathcal{O}.y) \lor max(\mathcal{O}.y) < s.y)$ then
            $satisfy \leftarrow$ False
return $satisfy, valid$
```

```
Algorithm 5: Test for same color relation
Input: all referents $\mathcal{A}$
Output: boolean $satisfy$, boolean $valid$
$valid \leftarrow |\mathcal{A}| > 1$
// Check if range of color is
   smaller than the threshold
$satisfy \leftarrow valid \land$
 $max(\mathcal{A}.color) - min(\mathcal{A}.color) < 30$
return $satisfy, valid$
```