The red cup on the left: reference, coreference and attention in visual dialogue

Simon Dobnik[1,2] and **Vera Silfversparre**[1]
[1]Department of Philosophy, Linguistics and Theory of Science
[2]Centre for Linguistic Theory and Studies in Probability (CLASP)
University of Gothenburg, Sweden
simon.dobnik@gu.se, gussilfve@student.gu.se

Abstract

We examine how conversational partners refer, co-refer and direct attention in conversations over a visual scene. Using an extension of the CoNLL annotation scheme for coreference for the visual domain we annotate the Swedish part of the Cups corpus. The annotation consists of identifying noun phrases and assigning them IDs of entities in the visual scene. We perform quantitative and qualitative linguistic analyses of the annotated data which point towards interesting observations of how conversational participants direct attention: it is likely that entities are co-referred to within the same conversational game, for spatial descriptions there is a preference for lateral dimensions over front and back and more attention is directed towards entities that are visually ambiguous or those that are part of the task. Overall, we demonstrate that referential attention is driven by both visual and conceptual, task-related information.

1 Introduction

Visual dialogue takes place in some visual context either physical or virtual. Conversational participants discuss the visual scene but they also relate it to their beliefs, desires and intentions as defined by the task they are engaged in. An important challenge for building visual dialogue systems is to model how such perceptual information interacts with higher level conceptual aspects of their information state in order to generate and interpret referring expressions such as "the red cup on the left" in the setting of a collaborative dialogue (Clark and Wilkes-Gibbs, 1986; Byron, 2003). In this paper we examine referring expressions in visual dialogue, in particular the mechanisms behind how speakers and hearers generate and interpret them in a highly visually and linguistically ambiguous environment. Conversational partners must rely on mechanisms of *attention* that assigns salience to contextual information from the visual, linguistic and task-related domains which affects how referring expressions are generated and interpreted (Kelleher et al., 2005). Literature on attention (Lavie et al., 2004) distinguishes between perceptual selection, a process that selects relevant visual features, and cognitive control, a process that selects the relevant conceptual information, which compete for the same cognitive resources. Since conversational participants are engaged in a collaborative task joint attention will be aimed at.

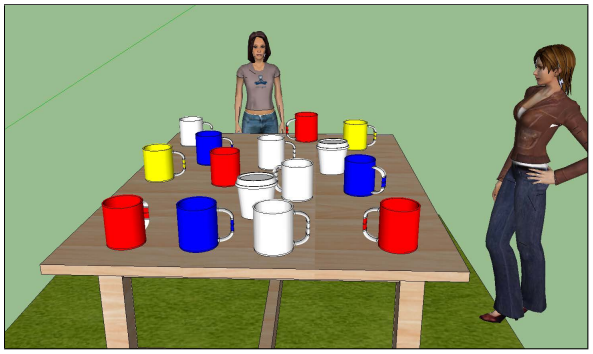
Reference and coreference resolution has been studied both in the domain of the textual documents (Sukthanker et al., 2020), in the domain of situated dialogue (Kelleher et al., 2005; Rolih, 2018; Smith et al., 2011) or in the domain of vision and language (Kottur et al., 2018; Yu et al., 2019). In the domain of textual coreference, one of the most known resources is the section of the OntoNotes corpus annotated for coreference as a part of the CoNLL-2011 shared task (Pradhan et al., 2011). Another well-known resource is the ARRAU corpus (Poesio et al., 2018; Uryupina et al., 2020). In the domain of visual dialogue the SCARE corpus (Stoia et al., 2008) contains spoken dialogues in a virtual reality maze environment with buttons, cabinets and doors. We follow the tradition of coreference annotation in the textual domain, by starting with the CoNLL 2011/2012 scheme and extending it to the domain of the visual dialogue of the Swedish part of the Cups corpus (Dobnik et al., 2015, 2020). The corpus is different from other corpora used in research on referring in that it comes with a single visual scene with a known ground truth representation of entities from which different views can be generated and over which participants can engage in long dialogues. This makes it an ideal candidate for studying coreference. Appendix A.1 shows some examples discussed here. Based on this annotation we address the following questions:

Proceedings of the 25th Workshop on the Semantics and Pragmatics of Dialogue, September 20–22, 2021, Potsdam / The Internet.

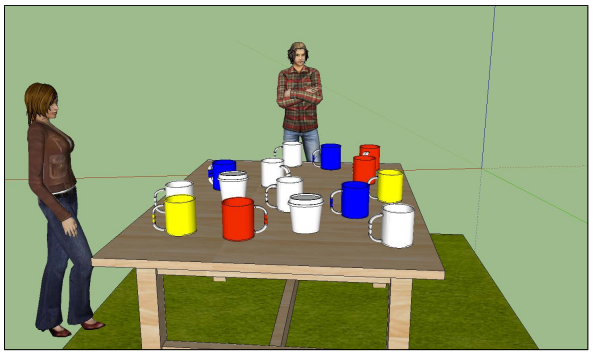
Q1: How do the interlocutors refer and co-refer to entities in the visual scene?

Q2: What are the issues with the referent annotation when starting with an annotation scheme developed for the textual domain and how can they be addressed?

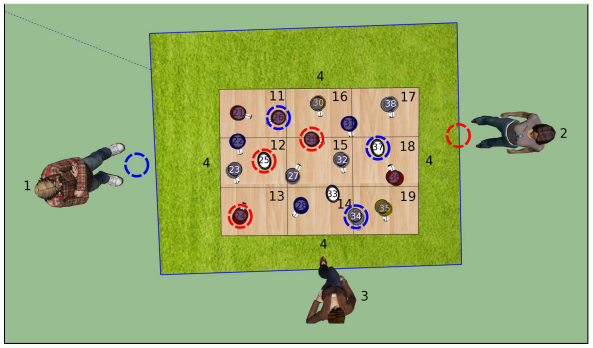
Q3: How is the attention (estimated from the reference of descriptions) distributed over the visual scene?



(a) The view of P1



(b) The view of P2



(c) Ground truth view of the scene

Figure 1: The scene as seen by P1 (a) and P2 (b). (c) shows a top-down view of the scene with all objects included and their object IDs. Objects marked with coloured circles cannot be seen by a participant marked with the same colour. P3 is a passive observer Katie.

2 Data and annotation

The Cups corpus (Dobnik et al., 2015, 2020) was created to examine collaborative dialogue over a visual scene and therefore resembles the Map Task (Anderson et al., 1991). It was previously used to study selection of reference frames, dialogue games (Storckenfeldt, 2018) and coreference (Dobnik and Loáiciga, 2019). A virtual scene containing a table with cups of different types and colours, two active conversational participants at the opposite sides of the table and a passive observer has been created in 3d-modelling software as shown in Figure 1. A static view of the scene was created for each participant. In addition, some objects were removed from the view of each participant but these were kept in the view of the other participant. The same views were used for all participant pairs. The data collection was done in a lab environment. Participants are instructed to interact over a textual computer interface in order to find and make a note of the missing cups which defined their collaborative task. To encourage spontaneous longer dialogue the task was not restricted in time. The nature of the task prevented participants from communicating through intonation, prosody, eye-gaze and body gestures. Table 1 summarises the current size of the corpus. We refer to the corpus as (sv.P05.100) which stands for the 100th turn of the P05 dyad of the Swedish sub-corpus.

Corpus	Dialogue	Turns	Native speakers of
English	en.P01	157	Swedish
	en.P02	441	English
Swedish	sv.P01	118	Swedish
	sv.P02	114	Swedish
	sv.P04	75	Swedish
	sv.P05	163	Swedish
	sv.P06	248	Swedish
	sv.P07	308	Swedish

Table 1: The Cups corpus. Here we annotate and analyse the Swedish (sv) part.

For the purposes of this study we annotated the Swedish sub-corpus (sv) with the CoNLL 2011/2012 annotation scheme (Pradhan et al., 2011) used for textual data but in contrast to OntoNotes (Pradhan et al., 2011) we annotate all noun phrases, as in the ARRAU corpus (Poesio et al., 2018). Note, however, that OntoNotes also contains annotation of coreference for verbs and temporal expressions. The annotation was done by the second author and then interesting and challenging examples were discussed with the first au-

Dlg	P01	P02	P04	P05	P06	P07	Total
RFs	197	360	278	395	463	571	2264

Table 2: The number of referring expressions in the Swedish part of the Cups corpus per dialogue.

thor. Based on this discussion, annotations were adjusted and notes were made for the annotation manual. The annotation file is automatically tokenised and then the annotation consists of two parts. First, noun phrases are identified using the BIO tags (`B-NP`, `I-NP` and `O`). Then, co-reference chains are identified over noun phrases by assigning referent IDs to them, e.g. `(11,13` for the opening word of a noun phrase and `11,13)` to the closing word of the same noun phrase while no tag is assigned to the intermediate words. In the standard textual coreference annotation (OntoNotes and ARRAU), the IDs are incremented as new referents are introduced in the text. However, in our work we pre-identify referents as entities (participants, objects and regions) identifiable in the visual scene as shown in Figure 1c. In this respect our annotation resembles the annotation of the SCARE corpus of visual dialogue (Stoia et al., 2008) where IDs are also pre-assigned to entities in the visual environment but is different from it in that we extend the assignment of IDs in two ways. For NPs that cannot be assigned a referent special tags were introduced and additional numeric tags were assigned to entities in the scene that were not previously identified in Figure 1c (see Section 3.2, (sv.P06.64-67)). Overall, in our adaptation of the CoNLL annotation scheme all noun phrases are annotated, a particular entity always has the same ID and a single noun phrase can be assigned several IDs. Referring expressions with the same IDs are coreferential. An example annotation is shown in Appendix A.2. All annotations are available at https://github.com/sdobnik/cups-corpus.

3 Results

3.1 Reference and coreference to entities (Q1)

Table 2 shows the number of referring expressions used in individual dialogues and in total in the Swedish part of the Cups corpus. The counts vary across different dialogues: for example P07 contains approximately three times as many referring expressions compared to P01. These referring expressions are assigned 3,867 references to entities in total which is 1.71 times the number of referring

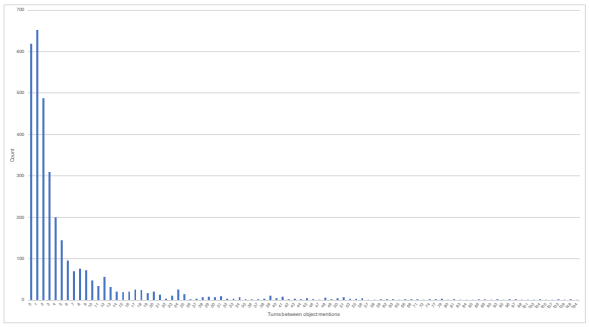


Figure 2: Latency of turns before an object is re-referred to. Latency 0 means that the objects is re-referred to in the same turn.

expressions. This indicates that very frequently an expression is referring to more than one entity. Out of 3,867 references to entities there are 3,515 references to entities with pre-defined IDs and 352 references to entities without IDs that were added dynamically by annotators. This shows that participants primarily refer to and attend to the entities that were pre-annotated in the scene which was done relative to the task and the goal of the conversation participants engage in. However, 352 (9.1%) of references to entities not pre-annotated indicates that the notion what is an entity (an object or a region) might not be straightforward. For example, the participants might refer to parts of the objects in order to disambiguate them, e.g. "sen en vit med lock, den står något närmare dig, sen en vit med handtag" (Then a white with a lid, it is somewhat closer to you, then a white with a handle.) (sv.P07.26-29). References to new objects are also frequently required when referring to regions as participants create regions and rows dynamically based on the topology of the scene and internal relations of objects rather than a global and equal geometric split of the scene.

Since conversational participants have access to the same visual scene throughout the dialogue objects or regions may be visually or linguistically attended more than once. To quantify how objects are re-referred to over the course of the dialogues we calculate the number of turns between two consecutive references to the same entity (≠mentions). Figure 2 shows that coreference ranges from 0 to 194 turns. It is very common for the object to be referred to in the same turn and then within 1 to 4 different turns. After that, the coreference to the same object decreases fast. For example, the same object is re-referred to over 20 turns less than 20 times, and 10 times over 30 turns. That coref-

erence is focused on a smaller number of turns indicates that participants collaboratively discuss and re-refer to objects until the ambiguity (both visual and conversational) is resolved (sv.P02.36-41, dialogue in Appendix A.1). The distribution of coreference also indicates that the objects might be re-referred to within the scope of the same conversational game. (Storckenfeldt, 2018, p.28) reports that the mean length of the annotated conversational games in this corpus range from 2.9 to 5.5 utterances which corresponds to the coreference figures reported here. Longer coreference could then be explained by the fact that entities are re-referred to in another conversational game. A possible reason to return to an entity is to use it as a landmark or a comparison when locating other entities. Once an entity is visually and linguistically grounded in the common ground it becomes a part of the shared knowledge and therefore a useful referential landmark (sv.P06.21-24). We expect that the usage of landmark entities also decays in time and landmarks that were more frequently referred to are preferred (Kelleher and Dobnik, 2020). As they are salient in the common ground and reference to them is not under discussion anymore, they only need to be referred to once as landmarks. This also explains a drop in frequencies after 4 turns.

3.2 Reference, coreference and visual dialogue (Q2)

In this section we examine questions related to annotating reference and co-reference in visual dialogue using the CoNLL 2011/2012 annotation scheme and suggest its required extensions.

The first question relates to the annotation of expression that are not referring to the corpus scenes and therefore cannot be assigned an object ID. Swedish also uses a demonstrative pronoun *det* as in "det finns" or "det är" which corresponds to English "there is" and "it is/they are" (sv.P06.127-129). Such pronouns are not referring and we annotate them as *expletives*. Conversational participants may refer to *entities outside the visual scene*, for example "in my picture" referring to a printed sheet of paper with a visual scene (sv.P01.61), or "byracka" referring to the other participant in a friendly derogatory way (sv.P06.4-6). Thirdly, there may be *non-referring expressions* that are used. These could be to abstract entities "in princip" (in principle, basically) or negated expressions "ingen lockmugg" (no cup with a lid) (sv.P04.51-56) but negated expressions are sometimes referential referring to an object that the other person previously referred to (sv.P04.52) or referring to an object of not being of that kind. Fourthly, *interrogative noun phrases* occuring in direct and indirect questions are also non-referential, e.g. "vad" (what) (sv.P05.145), "vilken farg" (what colour) (sv.P06.174-175). Table 3 shows the distributions of annotations of these categories in individual dialogues and in total. Expletive and non-referring expressions are most common but also note that there are considerable differences between different dialogues, e.g. there are 54 Non-R in P07 but only 4 in P04. This indicates that different conversational dyads might use different referring strategies.

Dialogue	Ext	Expl	Non-R	Wh-Q
P01	5	6	8	2
P02	6	21	13	6
P04	5	17	4	1
P05	2	25	27	7
P06	9	23	17	7
P07	13	29	54	6
Total	40	121	123	29

Table 3: The distribution of expressions not referring to objects IDs: external reference objects (Ext), expletive expressions (Expl), non-referring expressions (Non-R) and expressions used in wh-phrases (Wh-Q).

The second question relates to how to apply the existing annotation scheme on the data. Noun phrases can be complex containing embedded noun phrases of the form NP Relation NP, for example "de vita med handtag och utan lock" (the white one with handles and without lids) (sv.P04.40) and "en röd mugg på din vänsterkant" (a red cup to your left side) (sv.P05.57). Should "handles" and "your left side" also be annotated? Motivated by the research on spatial relations where two NPs are distinguished as Target and Landmark we decided to annotate each NP separately, provided that they are referring to distinctive objects and regions. However, sometimes this convention becomes hard to follow and this is related to whether the NPs are considered as referring to entities or properties of a single entity, e.g. "en röd mugg med lite rött på handtaget" (a red cup with some red on the handle) (sv.P04.3) where "some red on the handle" was annotated as as a single entity rather than two distinctive entities. This convention is different from (Stoia et al., 2008) in the SCARE corpus where embedded noun phrases such as "this cabinet on the right" are annotated as belonging to the umbrella

noun phrase, possibly because here the annotation is limited by fixed pre-defined entities.

Note that participants see a slightly different scene where some cups are missing from their view which means that it may happen that they associate a certain description with different objects/cups. In other words, there may be a miscommunication of reference but which is normally resolved through clarification in dialogue once participants discover there are inconsistencies in their information states. Errors in the way the objects are described or expressions interpreted might also happen (sv.P05.110-115). In such cases we annotated expressions as referring to objects relative to the information state of the utterance speaker, the object that they intend to refer to. In most cases this can be resolved from the visual context of the speaker but sometimes the annotators have to guess about the cognitive state of the speaker.

To answer the question how difficult it is to annotate coreference using this annotation scheme before starting the annotation of the Swedish dialogues we re-annotated the first 14 turns or 250 words of one of the English dialogues (en.P02) for which annotations already exist (although there the annotator might have used different strategies as described above). To measure the agreement on noun phrase identification we calculate the κ coefficient on the BIO tags (B-NP, I-NP and O) which results in $\kappa = 0.84$. Unfortunately, κ cannot be used for referent identification as each noun phrase might refer to one or several referents. To estimate agreement on referent identification we calculate a Sørensen–Dice coefficient that measures the overlap of the identified referents of each noun phrase. We then average all coefficients over all noun phrases. The average Sørensen–Dice coefficient is $\bar{DSC} = 0.70$. Overall, there is a good agreement on both annotation tasks.

3.3 Reference and attention (Q3)

Examining what referents are referred to in dialogue might tell us something about participants attention on the scene. Examining this might give us important preliminary insights about the strategies of reference resolution. Figure 3 shows global reference to entities over all dialogues.

There is a tendency that participants (1 and 2) refer to themselves or each other the most (not so much to the passive observer Katie, 3), followed by the objects (21–38) and then regions (11–19).

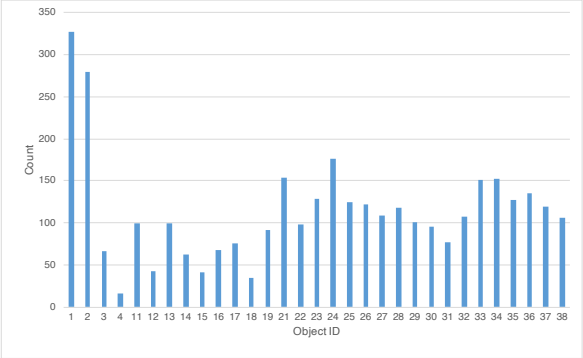


Figure 3: Reference to entities over all dialogues: 1-3 are participants, 4 is the table, 11–19 are regions and 21–38 are objects. See also Figure 1.

The participants' references to themselves reflect the collaborative nature of the task. Katie on the other hand is only used infrequently as a landmark to relate other objects to, for example: "På den sida där Katie inte står" (On the side where Katie is not standing) (sv.P02.48-49) or to set the spatial frame of reference or perspective on the scene "okej första raden fran katie pa hennes hogra sida. . . " (Okay, in the first row in front of Katie on her right side,. . .) (sv.P01.54-57).

Cups are more frequently referred to than regions which is expected as they are the objects of the task while regions refer to their locations. Note that 11, 13, 19 and 17 are the most attended regions. These represent corners of the table and are therefore good landmarks. Another reason why our pre-annotated and to participants invisible regions might be used less is that such geometric division is less natural for participants to refer to who dynamically create regions based on the object topology. For example, they do not say "mittenkvadraten närmast dig" (the central square closest to you) but "the second row closest to you" which does not match the geometric regions. We were aware of this when pre-annotating the grid and our hope was that the grid would provide some coarse granularity to annotate regions but in some cases it is hard to match the region referred to and the geometric region and in these cases labels for new dynamic regions were created. Sometimes it is hard to determine whether an expression is referring to a region, for example in elliptical noun phrases. We considered " din vänsterkant" (your left corner) (sv.P05.57.11-12), "till höger om den vita" (to the right of the white one) (sv.P05.39.10-14) as regions but "den står typ innanför den gula muggen" (It is roughly standing outside of the yellow cup)

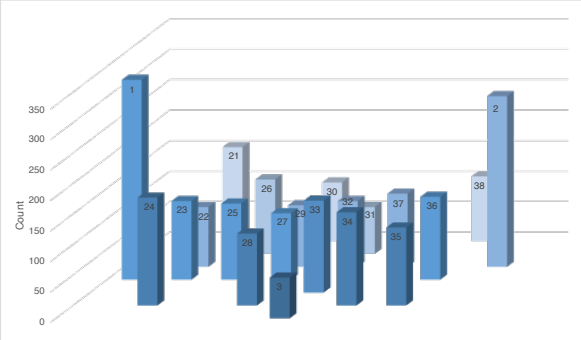


Figure 4: Attention over objects as measured by reference to them. The columns are arranged in the same spatial configuration as objects on the scene. Object 4, the table, is not shown.

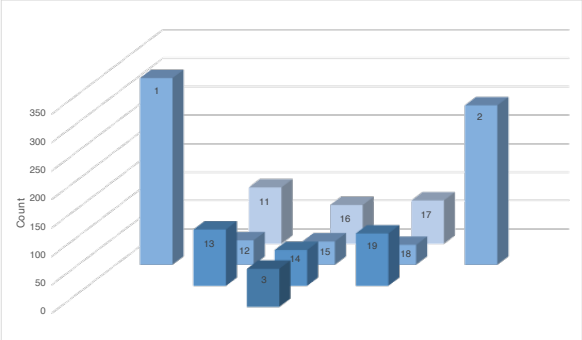


Figure 5: Attention over regions as measured by reference to them. The columns are arranged in the same spatial configuration as regions in the scene. For reference P1 (1), P2 (2) and Katie (3) are also included.

(sv.P04.12). This example demonstrates that the interpretation of these expressions as regions depends on the context.

Cups in the visually ambiguous configurations and cups that are missing from either participants view are referred to more often and therefore receive more attention. For example cup 24 which is hidden from P2 but also easily confused with cup 21 which is positioned in the opposite corner close to P1 (see Figure 1c). Moreover, cup 21 can also be confused with cup 26 which is close by and missing for P1. Similarly, there can be misunderstanding regarding cup 34 which is hidden from P1 but there is a similar cup 33 close by (sv.P06.24-26, dialogue in Appendix A.1) and cup 23 and 25 where the latter is hidden from P2.

Figure 4 shows the distribution of attention over the visual scene as measured by the reference to objects in dialogue. It can be seen that overall (with variations described previously) attention is more or less distributed over objects which can be explained by the nature of the task: the participants need to evaluate a consistency of each other's descriptions against the entire scene. There is a tendency that cups closer to Katie receive more attention than cups on the opposite side, the ranking being 24, 21, 34, 33... in descending order. Note that there is a similar ambiguity on both sides of the table: between 34 (not visible to P1) and 33 on Katie's side and 21, 26 (not visible for P1) and 29 on the opposite side. Therefore, there may be an effect of the presence of Katie on the grounds that she is an animate being and a good point of reference to relate other objects to (Lipp et al., 2004). However, she is not referred to specifically as she is not taking part in the task.

Figure 5 shows the distribution of attention over regions. Regions on the side of the table (13,14,19 and 11,16,17) attract more attention than regions in the middle (12,15,18). This indicates that participants prefer the lateral dimension over the front-back dimension when relating objects which coincides with observations from literature on spatial cognition: "från mitt håll står det en take-away bakom den vita muggen / snett vänster om" (From my perspective, there is a take-away behind the white cup. Diagonally to the left.) (sv.P05.37-44). Also, regions in the corners of the table (11,13,17,19) receive more attention than the middle regions on both sides (14,16), in fact these are also the most attended regions. This appears to be due to the fact that these corners are closest to participants who split the table in two halves: "mer på min sida än på din" (more on my side than on yours) (sv.P02.62-63). Note that closest to a participant does not mean closest to the speaker. Reference to participants is much higher than regions and so is reference to cups, presumably due to the nature of the task. Regions are mainly used as landmarks to describe the location of cups: "på kates vänstra sida innåt framför dig" (on Katie's left side in front of you) (sv.P06.58).

Comparing the attention over cups in Figure 4 with attention over regions in Figure 5 we can observe a low correspondence, e.g. attention on regions 13 and 11 might be associated with objects 24 and 21. However, when we sum the references to cups per regions, the cups that fall in the middle regions (12,15,18) are referred to more often than the cups that overlap with the lateral regions (11,16,17) and (13,14,19). Therefore, it could be that for the reference to the cups in the central regions the side

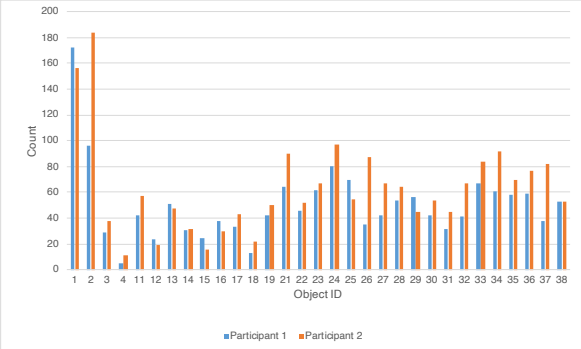


Figure 6: Reference to entities over all dialogues per participant: 1–3 are participants, 4 is the table, 11–19 are regions and 21–38 are objects. See also Figure 1.

regions are serving as landmarks, e.g. "Den står emellan den röda muggen på din vänstra sida och den gula muggen som står lite längre bort på din vänstra sida" (It is standing between the red cup on your left side and the yellow cup that is standing a bit further away on your left side.) (sv.P04.18).

Do conversational partners refer/attend to entities differently? Figure 6 shows reference to entities over all dialogues per participant. The reference to entities follows the same pattern for both participants which therefore also corresponds to the pattern in Figure 3. This shows that there is no preference for cups that would be closer and or more distant to a particular participant. Distance to an object does not seem to affect attention of that object. On the contrary, there is a tendency that objects that are hidden from the other participant (P1: 26, 34 and 37; P2: 24, 25 and 29) receive considerable attention. A likely explanation for this is that this is because conversational participants are engaged in a collaborative task that requires referring and subsequently co-referring to the same objects by the other partner until the task is completed (sv.P06.24-26, sv.P02.36-42, sv.P07.117-122, dialogues in Appendix A.1). Participant P1 refers to themselves more often than P2 and vice versa (sv.P06.220). (Dobnik et al., 2020) observe on the same dataset that the speaker's spatial perspective is used more often than hearer's. P2 refers to more objects than P1.

4 Discussion

We examined reference and coreference in visual dialogue. We argued that through patterns of reference in dialogue and the visual scene we can reconstruct patterns of attention that lead to production of these referring expressions. Due to the collaborative nature of dialogue these are also used by hearers to interpret referring expressions. Information how perceptual and discourse contexts interact in generation and interpretation of referring expressions is relevant for any computational application of vision and language as it allows us to resolve ambiguity that results through underspecification of referring expressions. As a starting point we took an established reference and coreference annotation scheme from the textual domain and adapted it to the domain of visual dialogue where linguistic expressions are also matched with referents grounded in the visual scene. This departs from the annotation strategies in the textual domain where discourse entities are introduced sequentially in text as they are referred to and then are subsequently re-referred to. However, in this domain discourse referents are already present once a participant sees and parses the scene: we indicated this by assigning participants, objects and regions fixed IDs. Additionally, we allow creation of dynamic entities and regions which are introduced in the same way as in the traditional textual co-reference annotation. Our notion of co-reference is also slightly different from the notion of co-reference in the text only domain. We do no specifically annotate coreference as a relation between referring expressions but this can be be inferred from the annotation scheme. We annotate a reference of referring expressions as a list of objects that an expression is referring to and hence if two referring expressions refer to the same objects then they are coreferential. Our annotation convention also allows us to compare referring expressions for partial (co)reference in case only some of the object IDs match. Additionally, object IDs could also be grouped and groups assigned IDs if coarse granularity of coreference would be required. For example, "(En 36) av (dem 21, 26, 36) står på (min sida 17, 18, 19), lite till (höger 18, 19) om (mitten 18)." (One of them stands on my side, a bit to the right of the middle) (sv.P02.19).

As conversation progresses, participants associate referring expressions with these entities based on how salient they are in the common ground; this is the reason why a description such as as "the red cup on the left" can be used successfully and the hearer can resolve its reference. We argue that the salience can be modelled as attention and the analysis of data in this paper is a first step towards computational modelling of reference and coref-

erence resolution in visual dialogue. Below we summarise our main findings.

Objects are most frequently co-referred to within the same conversational game. Our analysis of longer open dialogues shows that participants most frequently corefer to entities within 0 to 4 turns which coincides with the previous research on the length and structure of conversational games. These can also be nested. Conversational games depend on the collaboratory (sub)task that the participants are performing and once a task is complete and participants reach a mutual agreement they continue with a new task and a conversational game. Tasks are structured around a certain strategy which rarely considers the entire scene. Therefore, if a location of particular objects has been discussed, disambiguated and added to the common grounds of participants there is no need to discuss them again, unless they are used as salient landmarks for discussion of subsequent objects in new conversational games (sv.P02.73-82, dialogue in Appendix A.1). Structuring dialogue into sub-units explains why there are underspecified referring expressions since their scope can be resolved within the scope of these units.

The strategies to assign and resolve reference and co-reference are dynamic and creative. Although we have identified entities and regions in the visual scene within a certain level of granularity we frequently found cases where this was insufficient to fully capture the reference of the linguistic expressions. Firstly, not all noun phrases are referential, for example they can be expletives, referring to entities not present in the scene, non-referring (abstract and generics) or undetermined entities (wh-phrases and noun phrases used in questions). Here the challenge is that the same referring expression can be either referential or non-referential depending on the context in which it is used. For example, in "So maybe we could possibly go row by row, do you think? And say which cups are there? Or how should we work out where your unique cup are and vice versa?" does the speaker refer to specific alignments of cups, an abstract grid of rows or rows in general (en.P02.9)? Secondly, it is sometimes hard to decide what should be identified as a scene entity and what the granularity of regions should be, cf. our earlier example whether a sub-region of a handle constitutes a separate region (sv.P04.3). Thirdly, the same expression used by two conversational participants may be considered to refer to different entities by different conversational participants. These issues were resolved (i) by introducing four labels (expletive, external, non-referring and wh-questions) and annotation conventions (ii) by identifying referents on the basis of the information state of the speaker and (iii) by sometimes introducing new reference IDs to distinct parts of the visual scene dynamically. Even following these conventions, we sometimes had to make sub-optimal decisions in borderline cases. Overall, we were striving for regularity and consistency of the annotation scheme, so that it can be used in computational applications, as well as for informational richness and accuracy of semantic representation.

Reference of referring expressions points to spatial attentional patterns in the visual scene. For example, lateral regions are more attended than front-back regions while cups in the middle regions receive more attention. This is possibly because lateral dimensions serve as landmarks for describing target objects or cups or because front and back dimensions are referred to differently, relative to P1 and P2, e.g. "close to you" or as "left or right of Katie". Reference to lateral dimensions is frequently combined with the front and back dimension but this is described as a relation between objects rather than a reference to regions (and therefore may not be annotated): "sa till vanster; gul. sedan vit takeaway starx nedanför till höger…" (So to the left: yellow. Then a white take away just below to the right) (sv.P01.86-89). Note that spatial descriptions such as "vänster" (left) and "ovanför" (above) can either elliptically refer to regions or relations between objects. Distance of a participant to an object does not mean that a participant puts more attention to it. This is because participants collaboratively discuss all the objects in the scene, not just those that are close to them. Moreover, the data shows that certain cups receive attention when they are located in the areas where visual ambiguity is high, for example in regions where there are neighbouring similar cups or where there are cups hidden from a participant. Therefore attention to objects is driven by the task which is disambiguating the location of cups (sv.P05.64-88).

Overall, our work shows that conversational participants can communicate successfully in situations where both linguistic and visual information are underspecified. The underspecification is resolved from a variety of signals which do not

necessarily have a fixed meaning across all contexts. Strategies are chosen on the fly without a specific communicative signal which suggests that conversational participants need to reach agreements by processes of "virtual bargaining" (Misyak and Chater, 2014). This suggest that in visual co-reference resolution we should not look so much for patterns that can be directly extracted from the data as these might be context-specific but for communicative strategies that are available to participants and go beyond specific contexts. In (Loáiciga et al., 2021a,b) we compare reference and coreference in the Cups corpus with the Tell-me-more corpus (Ilinykh et al., 2019). The latter consists of shorter dialogues (normally one conversational game) over real images of environments that are different for each dialogue. The results indicate that the same strategies can be found in different contexts and tasks.

We examined the attention on the objects in the scene as a whole but it would also be important to examine how the attention changes when the dialogue unfolds. We expect that this would reveal interesting generalisations that would guide a computational system for co-reference resolution of individual referring expressions. A closer study of reference and segments of dialogues or dialogue games would also be in place as these may be natural boundaries of coreference. Are the same attentional patterns found within a dialogue and across the participant pairs? Is there a difference between English and Swedish dialogues? We have studied how referring expressions are mapped to objects but how are objects described by referring expressions (within conversational games) and how are referring expressions adapted when the same objects is re-referred to in the subsequent turns? Given the mechanisms of joint (visual and linguistic) attention how can linguistic forms be simplified and still be effective?

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. I am grateful to Sharid Loáiciga for important comments on various stages of this work. Three anonymous referees provided insightful and extremely useful comments of the original submitted version.

References

Anne H. Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, et al. 1991. The HCRC map task corpus. *Language and speech*, 34(4):351–366.

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.

Herbert H. Clark and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition*, 22(1):1–39.

Simon Dobnik, Christine Howes, and John D. Kelleher. 2015. Changing perspective: Local alignment of reference frames in dialogue. In *Proceedings of goDIAL – Semdial 2015: The 19th Workshop on the Semantics and Pragmatics of Dialogue*, pages 24–32, Gothenburg, Sweden.

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.

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.

Satwik Kottur, José MF Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. 2018. Visual coreference resolution in visual dialog using neural module networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 153–169.

Nilli Lavie, Aleksandra Hirst, Jan W de Fockert, and Essi Viding. 2004. Load theory of selective attention and cognitive control. *Journal of Experimental Psychology: General*, 133(3):339–354.

Ottmar V Lipp, Nazanin Derakshan, Allison M Waters, and Sandra Logies. 2004. Snakes and cats in the flower bed: fast detection is not specific to pictures of fear-relevant animals. *Emotion*, 4(3):233.

Sharid Loáiciga, Simon Dobnik, and David Schlangen. 2021a. Reference and coreference in situated dialogue. In *Proceedings of the Second Workshop on Advances in Language and Vision Research*, pages 39–44, Online. Association for Computational Linguistics.

Sharid Loáiciga, Simon Dobnik, and David Schlangen. 2021b. Reference and coreference in situated dialogue. In *Proceedings of the Second Workshop on Advances in Language and Vision Research*, pages 39–44, Online. Association for Computational Linguistics.

Jennifer B. Misyak and Nick Chater. 2014. Virtual bargaining: a theory of social decision-making. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1655):20130487.

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.

Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. 2011. CoNLL-2011 shared task: Modeling unrestricted coreference in OntoNotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–27, Portland, Oregon, USA. Association for Computational Linguistics.

Gabi Rolih. 2018. Applying coreference resolution for usage in dialog systems. Master's thesis in language technology, Department of Linguistics and Philology, Uppsala University, Uppsala, Sweden, June 14.

Cameron Smith, Nigel Crook, Simon Dobnik, Daniel Charlton, Johan Boye, Stephen Pulman, Raul Santos de la Camara, Markku Turunen, David Benyon, Jay Bradley, Björn Gambäck, Preben Hansen, Oli Mival, Nick Webb, and Marc Cavazza. 2011. Interaction strategies for an affective conversational agent. *Presence: Teleoperators and Virtual Environments*, 20(5):395–411.

Laura Stoia, Darla Magdalena Shockley, Donna K. Byron, and Eric Fosler-Lussier. 2008. SCARE: a situated corpus with annotated referring expressions. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pages 650–653, Marrakech, Morocco. European Language Resources Association (ELRA).

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.

Rhea Sukthanker, Soujanya Poria, Erik Cambria, and Ramkumar Thirunavukarasu. 2020. Anaphora and coreference resolution: A review. *Information Fusion*, 59:139–162.

Olga Uryupina, Ron Artstein, Antonella Bristot, Federica Cavicchio, Francesca Delogu, Kepa J. Rodriguez, and Massimo Poesio. 2020. Annotating a broad range of anaphoric phenomena, in a variety of genres: the ARRAU corpus. *Natural Language Engineering*, 26(1):95–128.

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 5122–5131, Hong Kong, China. Association for Computational Linguistics.

A Appendices

A.1 Referring in dialogue

(sv.P04.46-49)

46 P1: mellan den blå och gula$_{28,35}$, framför Katie, ser jag en mugg$_{33}$ med lock och utan handtag
Between the blue and yellow in front of Katie, I see a cup with a lid and without a handle.
47 P2: Står den$_{33}$ lite längre bort från Katie (lite mer mot mitten) än den gula$_{35}$ och den blå$_{28}$?
Is it standing a bit further away from Katie (a bit more towards the middle) than the yellow and the blue?
48 P1: lite mot mitten inte exakt mellan den blåa och gula$_{28,35}$
A bit towards the middle, not exactly between the blue and yellow.
49 P2: OK, den muggen$_{33}$ kan jag se.
Ok, I can see that cup.

(sv.P06.24-26)

24 P2: lite till vänster om den står en vit$_{34}$
A bit to the left of it, there is a white.
25 P1: ja som$_{33}$ har en annan form$_{25,33,37}$ och till vänster i hörnet en röd
Yes, that has another shape and to the left, in the corner, a red.
26 P2: jag har ingen röd! och den vita34 har samma form$_{21,22,23,24,26,27,28,29,30,31,32,34,35,36,38}$
I have no red! And the white has the same shape.

(sv.P02.36-42)

36 P2: Jag ser ju två röda$_{21,26}$ i ditt vänstra hörn
I see two red in your left corner.
...
40 P1: Ser du två röda$_{21,26}$ bredvid varandra$_{21,26}$?
Do you see two red next to each other?
41 P2: Precis, en$_{21}$ är längst ner i hörnet och en$_{26}$ är precis nedanför den$_{21}$ (från mitt håll sett)
Exactly, one is in the bottom corner and one is just below it (as seen from my perspective)
42 P1: Så är det inte för mig. Jag har en röd mugg$_{21,24}$ i varje hörn från där jag står.
It is not like that for me. I have a red cup in each corner from where I stand.

(sv.P07.117-122)

117 P1: den raden$_{21,22,23,24}$ som är närmast mig som du precis beskrev
The row closest to me that you just described.
118 P2: mm
Mm.
119 P1: bakom den$_{21,22,23,24}$ i din riktning står en take away-mugg$_{25}$
Behind it in your direction, there is a take away cup.
120 P1: ”på en ensam ””rad””$_{25}$”
On a separate row.
121 P2: ok, så rad två$_{25}$ för dig är en ensam take away mugg$_{25}$?
Ok, so row two for you is a solo take away cup?
122 P1: snett till vänster bakom den vita muggen$_{23}$ mitt framför mig
Diagonally to the left behind the white cup just in front of me.

(sv.P02.73-82)

73 P2: Sen är det två vita$_{33,34}$ kvar. En$_{33}$ har lock$_{5133}$ och en$_{34}$ har inte det. Den$_{34}$ som inte har lock$_{5125,5133,5137}$ står längst ut av dem$_{33,34}$, i princip framför Katie.
Then, there are two white left. One has a lid and one does not. The one that does not have a lid is positioned farthest out of them, basically in front of Katie.
74 P2: Ser du den$_{34}$ som står precis framför henne?
Do you see the one just in front of her?
75 P1: Nej det som står framför henne är en blå mugg$_{28}$.
No, what is in front of her is a blue cup.
76 P2: Hmm, okej. Finns det inget bakom den muggen$_{28}$?
Hum, okay. Is there nothing behind that cup?
77 P1: Den blåa muggen$_{28}$?
The blue cup?
...
79 P2: Precis. Ser du något bakom den$_{28}$?
Exactly. Do you see anything behind it?
80 P1: Nope
Nope.
81 P2: Okej... då tror jag att du kan anteckna att det står en vit mugg$_{34}$ utan lock$_{5125,5133,5137}$ precis framför Katie.
Okay... Then I think you can mark a white cup without a lid just in front of Katie.
82 P1: Okej, antecknat.
Okay, noted.

A.2 Annotation

Reference and coreference annotation of (sv.P04.25-26). The columns represent dialogue ID, participant ID, turn number, word number in turn, word, noun phrase tag, coreference annotation and English translation. The latter is only used here and is not part of the corpus annotation.

```
P04 1 25 1 jag    B-NP (1)         I
P04 1 25 2 ser    O                see
P04 1 25 3 tre    B-NP (22,28,31   three
P04 1 25 4 blåa   I-NP  22,28,31) blue

P04 2 26 1 Jag    B-NP (2)         I
P04 2 26 2 kan    O                can
P04 2 26 3 också  O                also
P04 2 26 4 se     O                see
P04 2 26 5 3      B-NP (22,28,31   3
P04 2 26 6 blå    I-NP             blue
P04 2 26 7 muggar I-NP  22,28,31)  cups
P04 2 26 8 .      O                .
```