Annotating anaphoric phenomena in situated dialogue

Sharid Loáiciga[1] **Simon Dobnik**[2] **David Schlangen**[1]
[1]Computational Linguistics, Department of Linguistics, University of Potsdam, Germany
[2]CLASP, Department of Philosophy, Linguistics and Theory of Science,
University of Gothenburg, Sweden
{loaicigasanchez, david.schlangen}@uni-potsdam.de,
simon.dobnik@gu.se

Abstract

In recent years several corpora have been developed for vision and language tasks. With this paper, we intend to start a discussion on the annotation of referential phenomena in situated dialogue. We argue that there is still significant room for corpora that increase the complexity of both visual and linguistic domains and which capture different varieties of perceptual and conversational contexts. In addition, a rich annotation scheme covering a broad range of referential phenomena and compatible with the textual task of coreference resolution is necessary in order to take the most advantage of these corpora. Consequently, there are several open questions regarding the semantics of reference and annotation, and the extent to which standard textual coreference accounts for the situated dialogue genre. Working with two corpora on situated dialogue, we present our extension to the *ARRAU* (Uryupina et al., 2020) annotation scheme in order to start this discussion.

1 Introduction

With the ease of combining representations from different modalities provided by neural networks, text and vision are coming together. There is a growing body of resources addressing a setting in which the visual context can be exploited to support a textual task, for example visual anaphora resolution. [1]

Several corpora have been developed in the domain of vision and language (V&L), for example corpora of image captions (Lin et al., 2014; Young et al., 2014; Krishna et al., 2017), images and paragraph descriptions (Krause et al., 2017), visual question answering (Antol et al., 2015), visual dialogue (Das et al., 2017) and embodied question answering (Das et al., 2018). Through these the V&L research has progressively moved from sentence descriptions to descriptions involving utterances and conversations, therefore adding complexity to their semantic representations. In parallel to the corpora, V&L systems have been developed but of course these are limited by the complexity of the task for which the dataset has been collected. The end goal of the current research is to move to a more complex linguistic setting involving multiparty dialogue and visual representations that go beyond individual images.

Anaphora resolution has been studied both in the textual and situated dialogue domains (cf. Sukthanker et al. (2020) for an extensive survey of anaphora and coreference; (Kelleher et al., 2005; Seo et al., 2017; Kottur et al., 2018; Yu et al., 2019; Dobnik and Loáiciga, 2019)). In the textual domain, this has been formulated as a standard task with several corpora annotated uniformly for the most part, while in situated dialogue each corpus presents its own individual solution (cf. (Kelleher et al., 2005; Smith et al., 2011; Pustejovsky and Krishnaswamy, 2020)). With the increasing interest in the combination of V&L in deep learning applications, multimodal resources are increasingly used in the context of traditional textual natural language processing (NLP) tasks. As such, it makes sense to consider a common annotation strategy both for the textual and situated dialogue domains, basing it on the rich work of textual anaphora resolution standards. Doing so, we also hope to get new insights about the semantics of reference in natural language.

Situated reference resolution involves grounding linguistic expressions in perceptual representations (Harnad, 1990) or representations of actions (Roy, 2005). Anaphora resolution, traditionally a textual task, involves linking linguistic expressions referring to the same discourse entities (Stede, 2012). While challenging, the task is defined by the familiar nature of written texts: linear, planned and structured; defining thus the mechanisms and devices

[1]Also known as coreference resolution in the NLP domain, here we follow Poesio (2016) in our terminology.

Proceedings of the 1st Workshop on Multimodal Semantic Representations (MMSR), pages 78–88

found in them. In resources combining V&L, however, the textual part is often a dialogue or pairs of question-answers. As a result, the coreference devices differ from those found in texts and are closer to actual conversations in which people create reference to entities on the fly. This of course comes with its own challenges, but there are also some relations made easier since they can be grounded in the image.

As V&L come together, there is therefore an increased need for extending resources for the task of visual anaphora resolution. This means engaging with the challenges along two axes:

- Dialogue: built by two speakers who each have their own mental state and cognitive process but who are communicating through referring expressions which are projected in the same conversation. As conversations are linear (one cannot go back to the past or to the future) linguistic coreference is linear.
- Shared physical context: simultaneous access to an image or other perceptual context. Same as in dialogue, the speakers have different viewpoints of the scene and need to build their individual mental states representing the scene guided by visual attention. However, once a representation of a visual scene is built, reference can be made to its representations in a non-linear fashion.

We present our extension to the ARRAU (Poesio, 2004; Artstein and Poesio, 2006; Uryupina et al., 2020) annotation scheme by analysing two situated dialogue corpora: the *Cups* corpus (Dobnik et al., 2020) and the *Tell-me-more* corpus (Ilinykh et al., 2019), shown below in Figures 1 and 2 respectively. This exercise proved useful to pinpoint in what ways the purely textual document scenario is different from the domain of embodied interaction both in terms of the semantics of interaction and annotation practices.

The *Cups* corpus contains a conversation between two participants over an (almost) identical visual scene involving a table and cups where participants have different locations. Some cups have been removed from each participant's view and they are instructed to discuss over a computer terminal in order to find the cups that each does not see. The ground truth of the visual scene is known as it has been artificially generated. It may take over an hour for the participants to solve the task and their activity results in free dialogue close to spoken conversations including phenomena such as clarifications, repairs, restarts and variable grammar. (The conversations are logged at a key-press level.) The *Tell-me-more* corpus consists of images accompanied with a short text of five complete sentences, collected by asking participants to describe the image to a friend, successively adding details in short constrained conversations. The genre of these texts is therefore mixed: in between standard text (as found in news text for example) and dialogue data which reflects the features found in conversations rather than written conventions.

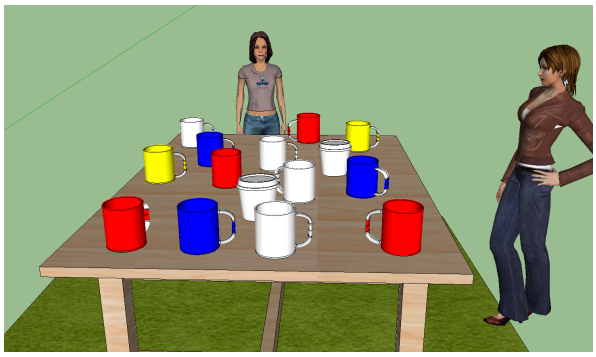
These corpora are complementary as *Cups* gives us accurate visual ground truth information with free and unrestricted dialogue, while *Tell-me-more* offers a richer unrestricted image with short and task-constrained (pseudo-)dialogues.

In this paper, we discuss a number of cases from these corpora that challenge both standard language grounding annotations as well as standard anaphora annotation. This work points thus towards required future work in creating anaphora annotation schemes that can handle situated dialogue.

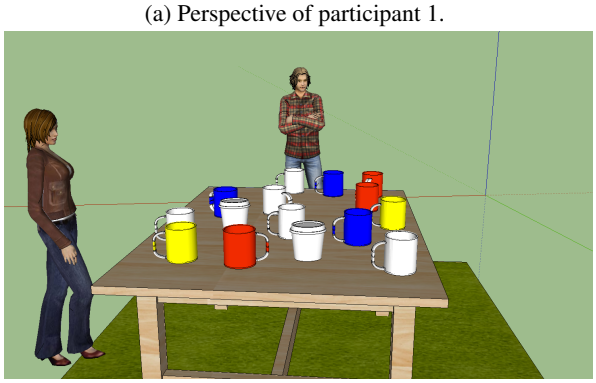
2 Related Work

Pointing to the inability of NLP tools to handle the textual part in situated dialogue, early works had described the need to ground the dialogue in the image in a manner informed by linguistics (Byron, 2003).

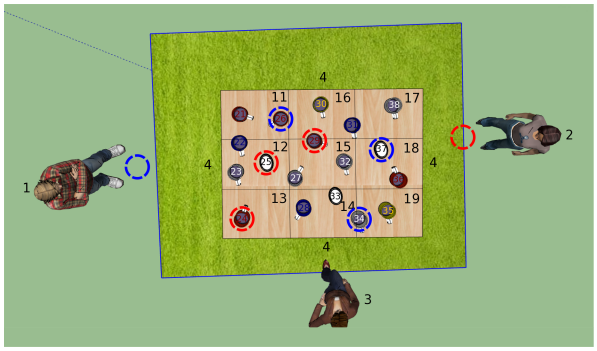
As content develops in a text, entities are introduced and re-mentioned, establishing discourse referents. The context is provided by the document and no extra-linguistic reference is needed for resolving the reference to an entity (Karttunen, 1969). In situated dialogue, on the other hand, the visual modality brings the extra-linguistic context as a source of referents. Here, resolving references to entities can be thus achieved by either looking at the picture or relating to the information that has been said previously in the discourse. Both of these processes happen simultaneously and therefore their interaction must be explained by theories of cognitive processing related to attention and memory (Kelleher and Dobnik). However, in order to understand both processes and their interaction we need to disentangle them. Extending the anaphora annotation paradigm is thus the best bet although not a lot of work exists in this area.



(a) Perspective of participant 1.

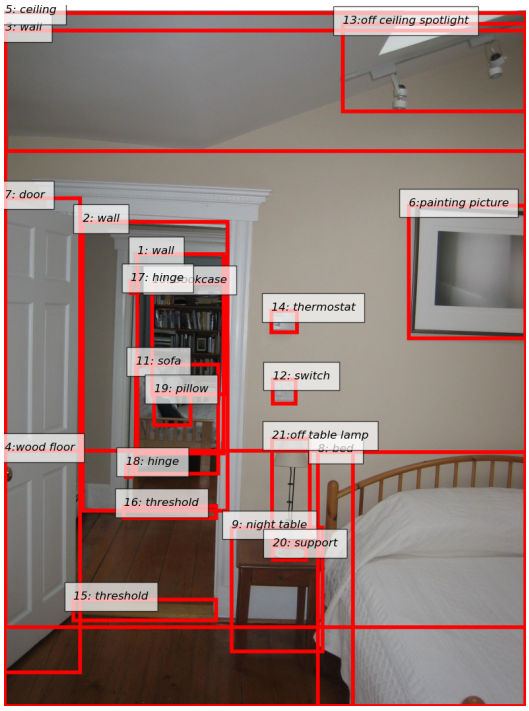


(b) Perspective of participant 2.



(c) Top-down perspective of the Cups corpus scene with ground truth object IDs.

Figure 1: Participant 1 cannot see the cups circled in blue, whereas participant 2 cannot see the cups circled in red. Person 3 is a passive observer of in the conversation.



1. it's a bedroom scene with the bed partially visible 2. the bed has a curved wooden headboard with slots like a fence 3. there is framed art hanging above the bed 4. to the left of the bed is a door, which is open 5. there is a small square nighstand next to the bed which has a lamp on top of it

Figure 2: Image and description sentences from the *Tell-me-more* corpus. Grammatical errors and other disfluencies are not corrected.

Textual coreference Annotated data for the coreference resolution task has mainly focused on news texts and concrete nouns, excluding reference to events and other coreferential relations such as bridging, deixis, and ambiguous items well documented in the linguistic literature but deemed infrequent or too difficult to process (Poesio, 2016). In contrast, there is a growing body of literature interested in phenomena beyond the nominal case (Kolhatkar et al., 2018; Nedoluzhko and Lapshinova-Koltunski, 2016), resulting in new annotated corpora (Lapshinova-Koltunski et al., 2018; Zeldes, 2017; Uryupina et al., 2020), although smaller in

size than OntoNotes (Pradhan et al., 2007), the largest and most used coreference corpus in the field.

Moreover, as a product of this year's edition of the CRAC[2] and CODI[3] workshops, a shared task on anaphora resolution in dialogues has been proposed. This will undoubtedly result in additional corpora annotated with the standards used for the coreference resolution task.

Visual coreference Coreference work based on the popular VisDial dataset (Das et al., 2017) targets only a limited set of referential expressions, partly because it relies on automatic tools (Kottur et al., 2018; Yu et al., 2019), which are known to be problematic with this genre. With a focus in grounded human interaction, there are corpora whose textual part comprises question answer pairs (Antol et al., 2015; Goyal et al., 2017). Those, however, are short in nature, with few opportunities for re-mention of the different objects in the image and hence coreference. Last, corpora designed towards navigation and location involve considerable dialogue interaction between instruction giver and instruction follower which include examples of coreference. For example, the SCARE corpus (Stoia et al., 2008) provides natural interactions, it has been audio recorded and then transcribed, the conversations are long and there are frequent referring expressions (it is hard to understand transcribed dialogues on its own), but overall the size of the corpus is small. Thomason et al. (2019) present a corpus of 2050 short human-human interactions in a virtual environment collected with crowd-sourcing.

Referring expressions generation The goal in this area is to generate referring expressions over several turns of conversation in a natural and non-repetitive way to the same (or different) grounded objects following principles of communicative discourse (Takmaz et al., 2020). Here, the PhotoBook dataset (Haber et al., 2019) is used. Our work is complementary to these approaches as it focuses on the interpretative rather than generative aspects of reference and coreference.

[2]Computational Models of Reference, Anaphora and Coreference, https://sites.google.com/view/crac2021/

[3]Workshop on Computational Approaches to Discourse, https://sites.google.com/view/codi-2021/accueil

3 The *ARRAU* annotation scheme

Deeply rooted in linguistic theory, the *ARRAU* corpus annotation scheme is particularly well-suited for annotating situated dialogue. Indeed, its annotation scheme was designed to accommodate different genres, including news, dialogue and narrative texts, and in consequence anaphoric phenomena beyond the nominal standard case typically found in other coreference corpora (Uryupina et al., 2020).

The dialogue genre has its own idiosyncrasies not covered by annotation schemes designed for news text, for example collaborative completions giving way for discontinuous markables (Uryupina et al., 2020), and more pronouns including deictics (Müller, 2007). The annotation scheme also includes guidelines for bridging reference, a much less studied type of reference but very commonly used in the *Tell-me-more* corpus discussed here. *ARRAU* is also known for containing annotations for both referring and non-referring expressions. Most coreference corpora focus on identity anaphora, meaning that only multiple mentions of the same discourse entity are annotated, leaving out those mentioned only once, also known as singletons. The large OntoNotes corpus, for instance, does not include annotations of singletons or expletives.

In the next section, we describe the general *ARRAU* annotation scheme along with our proposed adaptations. With the goal of moving towards general guidelines for the situated dialogue genre, the extensions we present target the common challenges of our two corpora.

4 Annotating situated dialogue

4.1 Mention identification and object detection

The first step is identifying the referring expressions or mentions to annotate. In *ARRAU*, all noun phrases are considered, marking the complete phrase with all its modifiers and not just its head. This includes noun phrases which are non-referring such as pleonastics and also noun phrases not re-mentioned later in the text. The mentions also include personal pronouns and demonstrative pronouns used as deictics (to refer back to non-nominal antecedents).

We also consider all noun phrases, including pronouns and deictics as mentions. For *Cups*, we created a simple NP chunker based on the regular

expression method (Bird et al., 2009) with moderate success: a manual annotation of one of the documents showed an error rate of about 30% (295 errors out of 1030 identified chunks). In contrast, for *Tell-me-more* we had annotators identify the NPs completely by hand.

Compared to *ARRAU*, the noun phrases in these corpora are rather simple, without a lot of modifiers. However, this does not mean that mention identification is straightforward as complex noun phrases with embedded markables such as *the blue cup with a white handle* do arrive. Consider also *the blue cup to the left of the red cup*, where a particular cup is referred to by taking another cup as a landmark: is it *the left* or *the red cup* or *the left of the red cup* which should be considered for re-mention?

Akin to the mention identification, the image in the multimodal corpora is processed in order to detect objects. In *Cups*, we have the ground truth of the scenes from which participants views have been generated. All the objects and geometrically defined regions are assigned a predefined ID as shown in Figure 1. In *Tell-me-more*, the object labels are part of the underlying ADE20K data (Zhou et al., 2017), extracted using tools from Schlangen (2019). Here, an automatic object classifier may not detect all the objects in the scene or assign them different labels than participants use when referring to them in the dialogue.

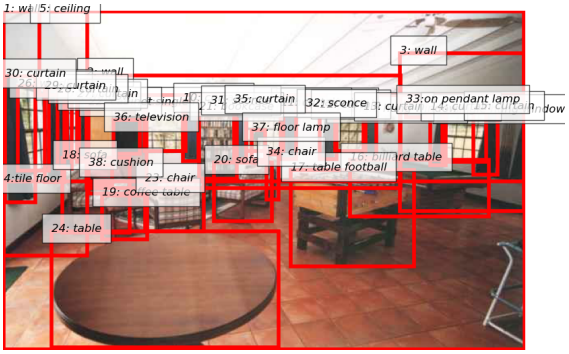
4.2 Characterisation of the mention

The morphosyntactic properties of the mention are annotated, including gender (female, male, neutre), number (singular, plural, mass) and person (1st, 2nd, 3rd), and its semantic type (person, animate, concrete, space, time, plan (for actions), abstract, or unknown). We include all these categories used in *ARRAU*.

In addition, we have also extended them in order to include a *cardinality* attribute. This accounts for a common strategy of grouping things in order to refer to them collectively. In other words, objects can be created dynamically as the dialogue progresses. For example, when a speaker refer to *the blue ones*, these are not all the blue cups in the scene but a particular set of blue objects that were grouped at that point of the dialogue and which can then be subsequently re-mentioned.

The *cardinality* attribute has the values *unique* and *group*. The first refers to objects represented by a single individual entity while groups refer to entities composed by several objects. Note that *group* is different from the *mass* number attribute in that mass nouns are usually singular. The value *group* refers to cases where the speaker decided to refer to a specific region of the image containing several entities together, for instance *green curtains* in sentence 4 in (1).

(1) 1. I see a picture of an entertainment room. 2. There is a round table in the foreground and a fussball table in the middle of the room, as well as a pool table further back. 3. There is a sitting area with chairs facing a television set. 4. The room has several windows with green curtains. 5. The floors are made of a brown tile.



4.3 Characterisation of the reference

As mentioned, *ARRAU* covers a broad range of anaphoric relations including both non-referring and referring noun phrases. Distinguishing between these two is non-trivial, and research around *ARRAU* have argued in favour of annotating both types (Poesio, 2016; Yu et al., 2020).

4.3.1 Non-referring

This includes mentions with a specific syntactic or semantic function: predication, expletive, idiom, incomplete or fragmentary expression, quantifier, and coordination. The last two are, by the authors own admission, controversial. Following *ARRAU*, we annotate all types of non-referential mentions.

4.3.2 Referring

If a mention is identified as referring, then its information status needs to be annotated as *discourse-new* or *discourse-old*; discourse-old information needs to point to an antecedent.[4] This distinction signals whether an entity is mentioned a first or subsequent time, shaping the reader's discourse model of that particular discourse entity (Stede, 2012).

[4] An antecedent can always be annotated as *ambiguous* if a clear entity cannot be identified for a particular mention.

Referring mentions yield coreference chains – the sequence of mentions pointing to a same entity in a text – a central construct in the coreference resolution domain. Built on top of the document as a unit, this notion relies on and in turn informs theories about accessibility hierarchy and salience of entities (Ariel, 1988, 2004; Grosz et al., 1995).

These theories are based on the observation that some forms are used to introduce entities and some others to refer to them: some entities are discourse-new and some are discourse-old. In situated dialogue, the image provides an additional context and source of referents, but it does not follow that the status of subsequent mentions is *old*. In the example (2) below, the fact that the discourse starts with *It* is licensed by the image and this source of reference should be accounted for differently in the annotation than a genuine discourse-old case such as the *it* in sentence 2.

(2) 1. It s a well-lit kitchen with stainded wooden cupboards. 2. There's a microwave mounted over the stove, which has a red tea kettle on it. 3. The appliances are black and stainless steel in the kitchen. 4. The countertops look like they're black granite. 5. The window has sunlight streaming in and it 's very brightly light.

In order to address these cases in the *Tell-me-more corpus*, we consider them discourse-old. Very importantly, in order to keep them distinct from genuinely *old* information in the discourse, we introduced a new value *task* for the antecedent (hence a discourse-old entity can have an antecedent which is a *phrase*, a *segment*, or the *task*). Our reasoning is that although the pronoun *It* does not have an antecedent in the text, it appears in the first position of the first sentence because the speaker was probably referring back to the *the image* in the instructions "Describe the image to a friend...".

In dialogue as found in the *Cups* corpus, on the other hand, references can be established either relative to utterances of a particular speaker or across utterances of different speakers, and in situated dialogue, references can also be established to the objects in the scene. This leads to another notable extension to the annotation scheme: the grounding of the entities to the image (Section 4.4).

4.3.3 Bridging

An understudied referential relationship also included in the *ARRAU* guidelines is bridging, i.e. an associative relationship between two mentions (Versley et al., 2016). When the status of a mention is either *new* or *old*, it is possible to annotate if the mention is a related object of some other entity. Here we follow the simplified scheme from Artstein and Poesio (2006):

- *Part:* "An object that stands in a part-of relation to an object previously mentioned."
- *Set:* "Relations that hold between a set and its elements, or between a set and a subset."
- *Other:* "Expressions containing the word *other* and referring to a second object of the same type as an object already mentioned."
- *Miscellaneous:* "Clear cases of bridging references that do not fall into any of the categories above."

The *Tell-me-more* corpus is rich in examples of bridging. Since the corpus uses pictures of different rooms in a house, after a room is introduced, typically a series of objects belonging to that room follow, creating many opportunities for using a bridging reference mechanism. For instance, image your surprise if the second sentence of example (3) started with *the toaster* instead of *the bed*. Coherence will be immediately broken.

(3) 1. This is a bedroom with a twin sized bed in it. 2. The bed has a blue bag laying on it and a green bad on the floor at the foot of the bed. 3. There is a nightstand aside of the bed with a water bottle on it. 4. There is an arched closet space on one wall and an arched shelving area too. 5. There is a small lamp attached to the wall at the head of the bed.

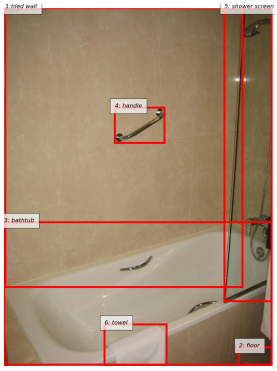
4.4 Grounding and referentiality

In spoken discourse people try their best to ground the references so they make sure they understand each other. To do so, they rely on the mechanisms of memory and attention (Kelleher and Dobnik). Memory controls how long objects referred to and objects perceived are cognitively salient in the mind of an agent, while attention controls the ratio of information that becomes salient coming from perception vs the amount of information coming from cognitive control of an agent (Lavie et al., 2004). Most entities annotated as concrete references can be grounded to the image easily. Following the *ARRAU-trains* annotation closely, we have added an attribute *on-image* with values *yes*/*no*. If the value is *yes*, then the atribute *bounding-box* with values *yes*/*no* needs to be annotated as well. The idea here is to distinguish between grounded entities detected by the object detector, and those that although visible do not have a bounding box or predefined ID.

This last scenario can be difficult, such as *base*

of the tub in example (4), where the object detector failed to recognise the target object. We observed, however, that this happens when the speakers refer to parts of the objects, and then the bridging annotation scheme can be smoothly applied.

(4) 1. This is a picture of a bathtub. 2. The tub is white. 3. The wall and base of the tub are brown. 4. The door appears to be glass. 5. There is a handrail on the side wall.



For bridging references, if a mention which is visible is in a *part-of* relation with another object which does have a bounding box, then we ground it to that object as well.

This process of referring to sub-objects is also fairly common in *Cups*. For example, participants refer to the cups handles and tops that we did not identify earlier.

Last, the image also allows for typically semantic properties to be used to refer back to the objects: colour, shapes, sizes. These can be genuinely referential (a form of ellipsis) or used in attributive manner. Compare for example *white* in the second sentence of (4), with (5) below.

(5) P1: closest to me, from left to right red, blue, white, red
P2: ok, on your side I only see red, blue, white

Note that in the case of mentions annotated as *groups*, we ground all the elements belonging to the group. However, deciding which elements exactly the speaker had in mind can be ambiguous. In (6) from Cups the speakers refer to *rows* of objects even though these are not arranged in strict geometric lines. Hence, what objects are included in a row is contextually defined and not always clear.

(6) P2: ok, so your next row
P2: you said there 's a takeaway cup somewhere marooned all alone
P1: Okay. So we have that row I described with the now found red cup. Then a takeaway cup that is between that row and the next. It's very much in the middle of the two rows.

Moreover, we observe references to different regions of the image, and these references change dynamically throughout the conversation, e.g. *my left*, *your right*, *the first row*. In the *Cups* corpus, we have split the scene into equal rectangular regions that are splitting the table into a grid as shown in Figure 1c. However, the grid nature of the subregions and their granularity are frequently insufficient as participants do not split the table to subregions in a grid-like manner but relative to the current focus on the scene and the topological arrangements of objects. In the example, "the empty space in the second row of objects close to you" an empty space has been designed as a new region which does not correspond to our projected grid-like regions. The references such dynamic objects must be resolved by the hearer and misunderstanding may occur, depending on the complexity and ambiguity of the scene.

Last, in the Cups corpus objects may be re-referred to again in different parts of a dialogue, potentially creating very long distance relationships between mentions. However, we generally restrict these to the scope of the dialogue games for which some parts of the corpus are also annotated.

4.5 The annotation process

Our annotation is implemented using the MMAX tool (Müller and Strube, 2006) for compatibility with the *ARRAU* MMAX schemes. An example of the annotator interface is presented in Figure 3. Besides the authors, three student assistants have been involved as annotators until now. We expect to release a first version of the annotation later during the year. This will include proper inter-annotator agreement metrics in order to evaluate the adequacy of the proposed schema.

4.6 Unaddressed challenge: speakers' cognitive state

Contrary to a Gricean-based analysis of spoken discourse, coherence-based theories of discourse do not traditionally take the cognitive state of the speaker as a necessary element to text interpretation (Bender and Lascarides, 2019). In situated dialogue, however, although the image can be treated as the ground truth of the situation, the speaker's cognitive state has to be considered by the hearer, in order to disambiguate the utterances. In other words, the hearer makes a model of the beliefs, desires and intentions associated with the utterance. This is exemplified in the following excerpt from

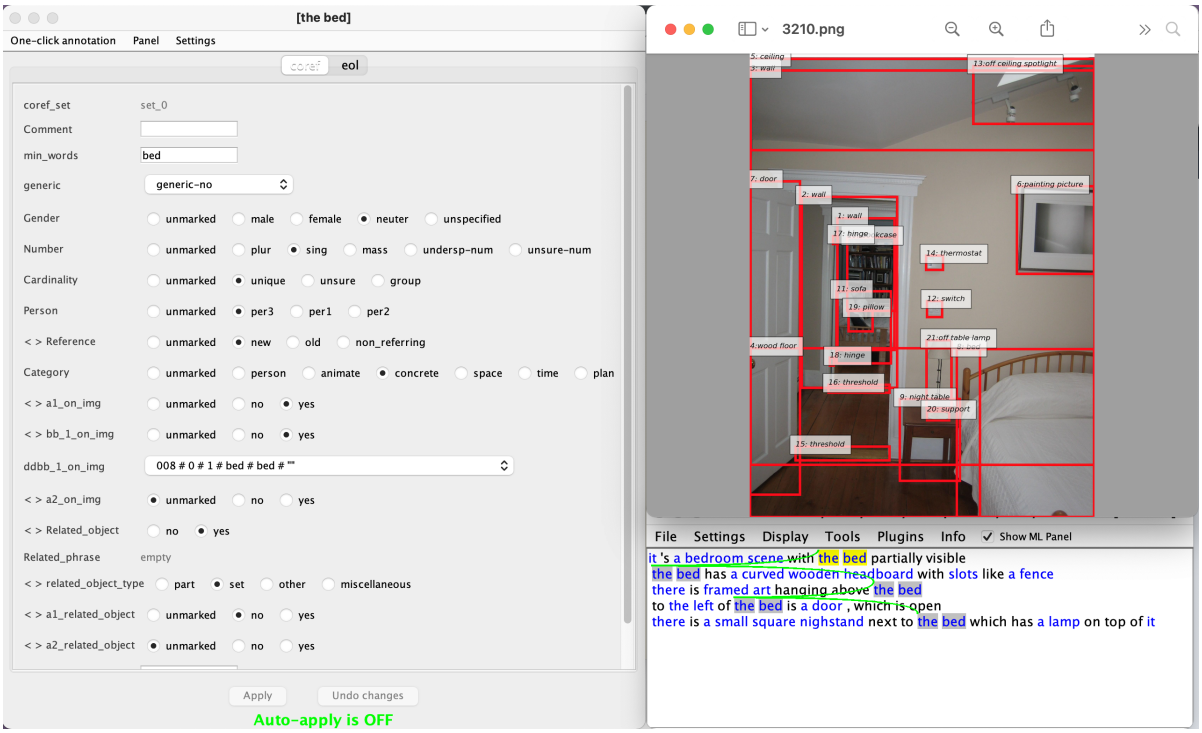


Figure 3: Example of annotation in the MMAX tool. Coreferential links are shown with the green lines in the bottom right. The annotator has simultaneous access to the image and the text while annotating all specified attributes in the annotation scheme.

Cups where both participants do not see one of the two red cups close by, but each a different one. They mistakenly believe that there is only one missing red cup and this dis-alignment of their beliefs gradually leads to increasingly diverging cognitive states.

(7) P2: there is an empty space on the table on the second row away from you
P2: between the red and white mug (from left to right)
P1: I have one thing there, a white funny top
P2: ok, i'll mark it.
DIALOGUE_STATE: B found O-25.
P1: and the red one is slightly close to you
P1: is that right?
P1: to my left from that red mug there is a yellow mug
P2: hm...
P2: can't see that and now i'm confused
DIALOGUE_STATE: B cannot see O-29.
P2: describe the second row away from you like you see it
P1: only one thing there, a white funny top
P2: aha, so it's closer to you than those i call "the second row"
P1: behind that, there is a yellow, red, white and blue
P1: from my left to right
P1: yes, that must be it!
P1: so what do you see in the "second row" from my perspective?
P2: i see a red, then space, then white and blue (same as katie's")
P2: no yellow
P2: is it on the edge of the table?
P2: on your left
P1: ok, yes!
DIALOGUE_STATE: inconsistent

5 Conclusions

Different V&L resources provide with an opportunity to explore the notion of discourse entity and (co)reference in grounded context. Since the nature of contexts defined by the tasks in which the corpora were collected varies considerably we get an opportunity to study the phenomena over these contexts and get a more complete picture of reference. Extending the coreference annotation to the V&L domain is essential to understand the relationship between reference and coreference. Work around textual coreference has defined the task with insufficient consideration of the semantic aspects involved in the interpretation of anaphoric phenomena; whereas work from the V&L community assumes that coreferential information can be inferred latently. By extending the coreference annotation scheme to rich situated dialogue corpora, we make explicit the relations at play between the text and the image. The same mechanisms that humans adopt to solve coreference in the textual domain should underlay results in the V&L domain.

Indeed, reference is underspecified in both modalities; any kind of information extraction from these domains will benefit from mechanisms that resolve this underspecification: capturing coreference is a door to capturing coherence. Furthermore, a rich annotation scheme that is portable between tasks and contexts, leads to the development of corpora allowing the training of data driven systems for the V&L domain and social robotics.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Mira Ariel. 1988. Referring and accessibility. *Journal of Linguistics*, 24(1):65–87.

Mira Ariel. 2004. Accessibility marking: Discourse functions, discourse profiles, and processing cues. *Discourse Processes*, 37(2):91–116.

Ron Artstein and Massimo Poesio. 2006. Arrau annotation manual (trains dialogues).

Emily M. Bender and Alex Lascarides. 2019. Linguistic Fundamentals for Natural Language Processing II: 100 Essentials from Semantics and Pragmatics. *Synthesis Lectures on Human Language Technologies*, 12(3):1–268.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python*, 1st ed edition. O'Reilly, Beijing, Cambridge, Farnham, Köln, Sebastopol and Tokyo.

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.

Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. 2018. Embodied question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2054–2063.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 326–335.

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 the 23rd Workshop on the Semantics and Pragmatics of Dialogue - Poster Abstracts*, London, United Kingdom. SEMDIAL.

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 *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A framework for modelling the local coherence of discourse. *Computational Linguistics*, 2(21):203–225.

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.

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

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.

Lauri Karttunen. 1969. Discourse referents. In *International Conference on Computational Linguistics COLING 1969: Preprint No. 70*, Sånga Säby, Sweden.

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

John D. Kelleher and Simon Dobnik. 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*, pages 41–50.

Varada Kolhatkar, Adam Roussel, Stefanie Dipper, and Heike Zinsmeister. 2018. Anaphora with non-nominal antecedents in computational linguistics: a survey. *Computational Linguistics*, 44(3):547–612.

Satwik Kottur, José 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)*.

Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. 2017. A hierarchical approach for generating descriptive image paragraphs. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3337–3345.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. 2017. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73.

Ekaterina Lapshinova-Koltunski, Christian Hardmeier, and Pauline Krielke. 2018. ParCorFull: a parallel corpus annotated with full coreference. In *Proceedings of 11th Language Resources and Evaluation Conference*, pages 423–428, Miyazaki, Japan. European Language Resources Association (ELRA). To appear.

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.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham. Springer International Publishing.

Christoph Müller. 2007. Resolving it, this, and that in unrestricted multi-party dialog. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 816–823, Prague, Czech Republic. Association for Computational Linguistics.

Christoph Müller and Michael Strube. 2006. Multi-level annotation of linguistic data with MMAX2. In Sabine Braun, Kurt Kohn, and Joybrato Mukherjee, editors, *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, pages 197–214. Peter Lang, Frankfurt a.M., Germany.

Anna Nedoluzhko and Ekaterina Lapshinova-Koltunski. 2016. Abstract coreference in a multilingual perspective: a view on czech and german. In *Proceedings of the Workshop on Coreference Resolution Beyond OntoNotes*, CORBON 2016, pages 47–52, Ann Arbor, Michigan. Association for Computational Linguistics.

Massimo Poesio. 2004. Discourse annotation and semantic annotation in the GNOME corpus. In *Proceedings of the Workshop on Discourse Annotation*, pages 72–79, Barcelona, Spain. Association for Computational Linguistics.

Massimo Poesio. 2016. Linguistic and cognitive evidence about anaphora. In Massimo Poesio, Roland Stuckardt, and Yannick Versley, editors, *Anaphora Resolution: Algorithms, Resources and Applications*, pages 23–54. Springer-Verlag, Berlin Heidelberg.

Sameer S. Pradhan, Lance Ramshaw, Ralph Weischedel, and Jessica MacBrideand Linnea Micciulla. 2007. Unrestricted coreference: Identifying entities and events in OntoNotes. In *International Conference on Semantic Computing (ICSC 2007)*, pages 446–453.

James Pustejovsky and Nikhil Krishnaswamy. 2020. Situated meaning in multimodal dialogue: Human-robot and human-computer interactions. Journal article manuscript, Department of Computer Science, Brandeis University.

Deb Roy. 2005. Semiotic schemas: a framework for grounding language in action and perception. *Artificial Intelligence*, 167(1-2):170–205.

David Schlangen. 2019. Natural language semantics with pictures: Some language & vision datasets and potential uses for computational semantics. In *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, pages 283–294, Gothenburg, Sweden. Association for Computational Linguistics.

Paul Hongsuck Seo, Andreas Lehrmann, Bohyung Han, and Leonid Sigal. 2017. Visual reference resolution using attention memory for visual dialog. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

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.

Manfred Stede. 2012. *Disourse Processing*. Morgan and Claypool Publishers, Toronto.

Laura Stoia, Darla Magdalene 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)*, Marrakech, Morocco. European Language Resources Association (ELRA).

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

Ece Takmaz, Mario Giulianelli, Sandro Pezzelle, Arabella Sinclair, and Raquel Fernández. 2020. Refer, Reuse, Reduce: Generating Subsequent References in Visual and Conversational Contexts. In *Proceedings of the 2020 Conference on Empirical Methods*

in Natural Language Processing (EMNLP), pages 4350–4368, Online. Association for Computational Linguistics.

Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer. 2019. Vision-and-dialog navigation. In *Conference on Robot Learning (CoRL)*.

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

Yannick Versley, Massimo Poesio, and Simone Ponzetto. 2016. Using lexical and encyclopedic knowledge. In Massimo Poesio, Roland Stuckardt, and Yannick Versley, editors, *Anaphora Resolution: Algorithms, Resources and Applications*, pages 397–429. Springer-Verlag, Berlin Heidelberg.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Juntao Yu, Alexandra Uma, and Massimo Poesio. 2020. A cluster ranking model for full anaphora resolution. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 11–20, Marseille, France. European Language Resources Association.

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.

Amir Zeldes. 2017. The GUM corpus: creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

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