Reference and coreference 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, a large number of corpora have been developed for vision and language tasks. 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. Working with two corpora approaching this goal, we present a linguistic perspective on some of the challenges in creating and extending resources combining language and vision while preserving continuity with the existing best practices in the area of coreference annotation.

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 coreference resolution.

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 multi-party dialogue and visual representations that go beyond individual images.

Situated reference resolution involves grounding linguistic expressions in perceptual representations (Harnad, 1990). Coreference 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 coreference mechanisms and devices 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 considerably from those found in texts and are closer to actual conversations whereby 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 coreference 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.
- Shared physical context: simultaneous access to an image or other perceptual context which enables non-linear references to it. Instead, the reference is guided by visual attention.

We present a linguistic perspective on these challenges by analysing a pilot annotation of two situated dialogue corpora: the *Cups* corpus (Dobnik et al., 2020) and the *Tell-me-more* corpus (Ilinykh et al., 2019), shown below in Figure 1 and example (1) respectively. Starting from the annotation scheme for several textual coreference datasets (Artstein and Poesio, 2006; Pradhan et al., 2007; Uryupina et al., 2019), this exercise proved useful to pinpoint in what ways the purely textual doc-

ument scenario is different from the domain of embodied interaction.

The first corpus contains a conversation between two participants over an almost identical visual scene involving a table and cups where participants have different locations (Figure 1). 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 *Tell-me-more* corpus consists of images accompanied with a small text of five complete sentences, collected by asking participants to describe the image to a friend, successively adding details. 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.

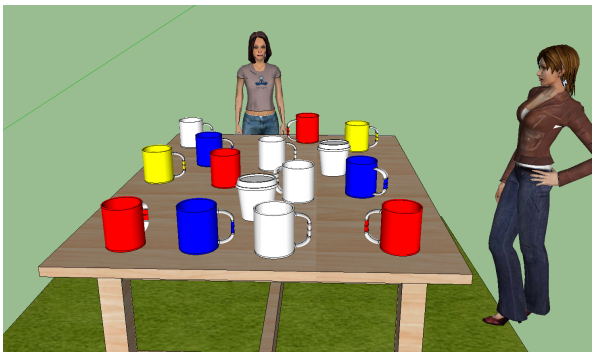
These corpora are complementary as *Cups* gives us accurate visual ground truth information with free and unrestricted dialogue, while *Tell-me-more* offers a richer unrestricted image with short and task constrained (pseudo-)dialogues.

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

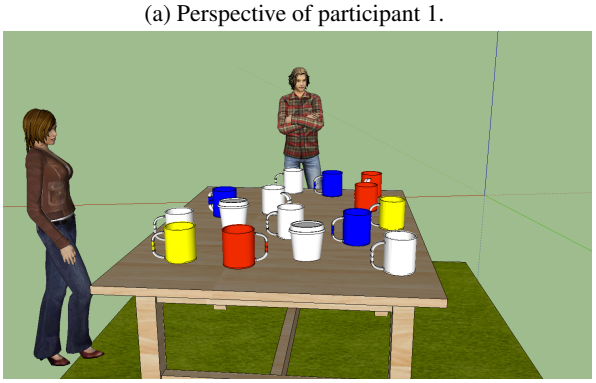
2 Related Work

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).

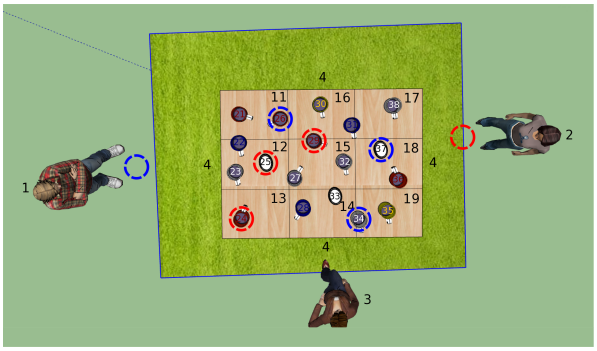
As content develops in a text, entities are introduced and re-mentioned, establishing discourse referents. The context is provided by the document and no extra-linguistic reference is needed for resolving the reference to an entity (Karttunen, 1969). In situated dialogue, on the other hand, the visual 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 by reading the discourse. Recording both strategies separately is crucial if we want to understand and model them soundly, in keeping with theories of cognitive processing (cf. (Kelleher et al., 2005)). Extending the coreference annotation paradigm is thus the best bet although not a lot



(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 visible to both participants as a reference point.

of work exists in this area.

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,

although still small in size, annotated corpora (Lapshinova-Koltunski et al., 2018; Zeldes, 2017; Uryupina et al., 2020).

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 (Stoia et al., 2008; Thomason et al., 2019) focusing on different kind of task and descriptions might be good candidates that could be explored and extended in a similar fashion as our corpora.

Referring expressions generation The goal in this area is to generate expressions over several turns of conversation in a natural and non-repetitive way, following principles of communicative discourse as for example in the recent PhotoBook dataset (Takmaz et al., 2020). Our work is complementary to such undertakings as it focuses on the interpretative rather than generative part.

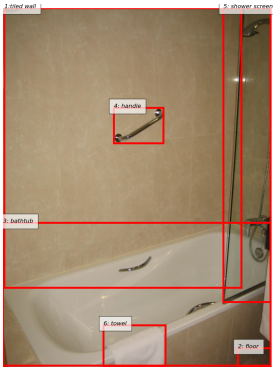
3 Understanding reference in situated dialogue

The notion of coreference chain–the sequence of mentions pointing to a same entity in a text–is central in coreference resolution. 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). In dialogue, however, references crisscross between the speakers and, one step further, in situated dialogue references crisscross between the speakers and the objects in the image. In this section we revise the annotation challenges in the annotation of anaphoric phenomena in data of this genre.

3.1 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 attention (Lavie et al., 2004). Although most concrete references can be grounded to the image easily, there are also some difficult cases. References can be found to portions of the image without a bounding box, such as *base of the tub* in example (1).

(1)



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.

In the previous example the difficulty arises because the object detector failed to recognise the target object. However, referring expressions are referential to a different degree, e.g., "Where are your blue ones?" – is the speaker referring to a particular subset of blue cups, all the blue cups in the scene, blue cups in general, or not referring to any particular set of objects? The distinction is sometimes not clear.

Last, as the image determines the scope of the referentiality, typical semantic properties are frequently used to refer back to the objects in the image: colour, shapes, sizes. These can be genuinely referential (a form of ellipsis) or used in an attributive manner. Compare for example *white* in the second sentence of (1), with (2) below.

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

3.2 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 to disambiguate their utterances, the hearer makes a model of their beliefs, desires and intentions associated with the utterance. This is exemplified in the following excerpt from *Cups* where both participants do not see one of the two red cups close by, but each a differ-

ent 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.

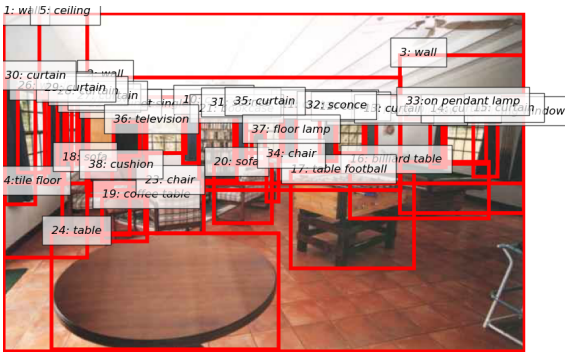
(3)
P2: there is an empty space on the table on the second row away from you
P2: between the red and white mug (from left to right)
P1: I have one thing there, a white funny top
P2: ok, i'll mark it.
DIALOGUE_STATE: B found O-25.
P1: and the red one is slightly close to you
P1: is that right?
P1: to 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

3.3 Level of specification

We observe a common strategy of grouping things in order to refer to them collectively. This raises the question: What is the level of specification needed in a coreference annotation? One could think about this in linguistic terms, for instance mass nouns or sets; alternatively, in computer vision, there is the distinction between things and stuff (Caesar et al., 2018).

In (4) below, is the reference to the *curtains* a case of a set composed by individual instances, or is it a mass noun? Note the curtains is a type of stuff in Caesar et al.'s work.

(4)
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.



In (5) from Cups, on the other hand, the speakers refer to *rows* of objects even though these are not arranged in strict geometric lines. Hence, which objects are included is contextually defined and not always clear.

(5)
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.

3.4 Information status

Different referring expressions have different properties and behaviour, an idea behind theories of salience and accessibility. They 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 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.

(6) 1. It s a well-lit kitchen with stainded [sic] 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.

4 Conclusions

V&L resources provide a unique opportunity to explore the notion of discourse entity in grounded context. Extending the coreference annotation to this domain is essential to understand the relationship between reference and coreference. 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 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.

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.

Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. 2018. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

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.

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.

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.

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.

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. 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.

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).

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.

Olga Uryupina, Ron Artstein, Antonella Bristot, Federica Cavicchio, Francesca Delogu, Kepa J. Rodriguez, and Massimo Poesio. 2019. Annotating a broad range of anaphoric phenomena, in a variety of genres: the ARRAU corpus. *Natural Language Engineering*, 26(1):95–128.

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.

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.