Anaphoric Phenomena in Situated Dialog: A First Round of Annotations

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

Abstract

We present a first release of 500 documents from the multimodal corpus *Tell-me-more* (Ilinykh et al., 2019) annotated with coreference information according to the ARRAU guidelines (Poesio et al., 2021). The corpus consists of images and short texts of five sentences. We describe the annotation process and present the adaptations to the original guidelines in order to account for the challenges of grounding the annotations to the image. 50 documents from the 500 available are annotated by two people and used to estimate inter-annotator agreement (IAA) relying on Krippendorff's α.

1 Introduction

Coreference resolution—linking together all referring expressions that refer to the same discourse entity—has a long tradition in computational linguistics. The progress is undeniable as evidenced by recent systems (Joshi et al., 2019; Kirstain et al., 2021), particularly in the text domain, rich in news data. Coreference resolution work for dialog and spoken data in general, on the other hand, has been less predominant, as the phenomena in this genre are broader and harder to tackle (Khosla et al., 2021). However, interest in tackling these challenges is possible thanks to the creation of new resources. Situated dialog corpora—dialog about an image presented as common ground to the dialog participants—is part of these recent resources. Dialog text approximates natural conversations, while the image constraints an otherwise unlimited choice of entities and events in the dialog.

In this paper, we present a first release of a portion of the multimodal corpus *Tell-me-more* (Ilinykh et al., 2019) annotated with coreference information according to the ARRAU guidelines (Poesio et al., 2021). The *Tell-me-more* corpus consists of images accompanied with a short English text of five complete sentences, collected by asking participants to describe the image to a friend. An example is presented in (1). The genre of these texts is therefore in between standard text (as found in news text for example) and dialog data which reflects the features found in conversations rather than written conventions. The simplicity of the text is essential for our purposes, as it allows us to test the limits of the guidelines to account for reference and grounding. In contrast, standard situated dialog is very rich in changes of point of reference, spacial references, and dynamic references depending on the participant's cognitive state that are very challenging to ground to the image.

(1) 1. There is four chair red laquer dining set shown in the image. 2. There are opened white french doors leading to the outside showing. 3. There is a pool with blue water showing through the french doors. 4. The pools is surrounded by green shrubbery. 5. The wood floor is covered with white paint.[1]

We discuss some of the changes to the baseline guidelines necessary to account for the challenges of grounding the annotations while following standard anaphora annotation. This release comprises 500 documents. From those, 50 documents are double annotated and used to estimate inter-annotator agreement (IAA).[2]

[1]Note that the examples have been transcribed with their original spelling errors and disfluencies. The English speakers who provided the data were recruited through Amazon Mechanical Turk and their IP addresses limited to the US.

[2]The annotations are publicly available at https://

Proceedings of the Fifth Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2022) , pages 31–37
Gyeongju, Republic of Korea, October 16–17, 2022.

2 Related Work

Anaphora resolution for situated dialog is a relatively unexplored area, reflected in the few resources available for it. The insufficiency of corpora hinders the learning from gold data which is standard in machine learning and has driven researchers to propose alternative strategies. Working with the VisDial dataset (Das et al., 2017), Kottur et al. (2018) use automatic coreference links generated with an out-of-the-box system[3], while Yu et al. (2019) annotate 5000 documents using workers recruited through crowd-sourcing. Li and Moens (2021), on their part, propose an unsupervised approach relying on heuristics by adding POS tags embeddings and sentence position embeddings in order to guide the system into learning noun antecedents. Note that these three papers deal with pronouns only, since they are frequent in the dialog genre.

Liu and Hockenmaier (2019) and Plummer et al. (2017), on their part, propose automatic methods to ground the entities in the text to specific regions in the image.

There exist other corpora whose textual part comprises question answer pairs (Antol et al., 2015; Goyal et al., 2017). Unlike dialog data, question answer pairs are short, with few opportunities for re-mention of the different objects in the image and hence coreference. There is also corpora designed towards navigation and location involving videos and long dialog interactions between an instruction giver and an instruction follower. Examples include the SCARE corpus (Stoia et al., 2008) and the corpus by Thomason et al. (2019). On a similar venue, recent work has used Minecraft[4] to collect dialog where an architect gives instructions to a builder about how to move and position some pieces in order to achieve a target structure (Narayan-Chen et al., 2019; Jayannavar et al., 2020). Due to the multiple changes in reference perspective and very long dialog games, this type of corpora is more difficult to annotate than the corpus used in this paper. In this sense, we see our work as a stepping stone towards achieving the annotation of more complex data in the future.

doi.org/10.5281/zenodo.7084861

[3] https://github.com/huggingface/neuralcoref

[4] https://www.minecraft.net/en-us

Average	Annotator A	Annotator B
tokens	48.16	48.16
mentions	13.72	17.08
singletons	9.38	12.4
chains	1.74	1.84
non-referring	1.86	2.02
bridging	2.64	3.4

Table 1: Annotators statistics averaged over 50 documents. We consider each set of 5 sentences a document.

3 The Annotation Process

The annotation was carried out by two annotators with a background in computational linguistics. The MMAX annotation tool (Müller and Strube, 2006) was chosen with the aim to replicate the ARRAU scheme easily.

3.1 Markables

Text Mentions. Annotators start by identifying the referring expressions or mentions to annotate. Following ARRAU, we consider all noun phrases (NPs) and instruct annotators to mark the complete NP with all its modifiers and not just its head. This includes NPs which are non-referring such as pleonastic NPs and also NPs not re-mentioned later in the text (singletons). The mentions also include personal pronouns and demonstrative pronouns used as deictics (to refer back to non-nominal antecedents).

Unlike ARRAU, the mention identification process is done entirely by hand. The absence of automatic preprocessing to detect the mentions resulted in a different number of mentions per annotator, as shown in Table 1. In addition, the annotators had a relatively high disagreement rate on the mentions boundaries, but not on the overall number of mentions, as the documents are short and simple. We analyze these disagreements further in Section 4.

Image Objects. The image, on its part, is processed automatically in order to detect objects and mark them with bounding boxes. In *Tell-me-more*, the object labels are part of the underlying ADE20K data (Zhou et al., 2017), extracted using tools from Schlangen (2019).

Mention Attributes. The morphosyntactic properties of the mention are annotated, including gen-

der (female, male, neutre)[5], 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.

An additional attribute of our own is *cardinality*. This accounts for a common strategy consisting on grouping things in order to refer to them collectively. In other words, objects can be created dynamically as the dialog progresses. The *cardinality* attribute has the values *unique* and *group*. The first refers to single individual objects while groups refer to entities composed by several objects. The value *group* is used for cases where the speaker refers to a specific region of the image containing several entities together, for example, *a four chair red laquer [sic] dinning set* in example (1) which is grammatically singular but conceptually plural.

3.2 Reference

As mentioned, ARRAU covers a broad range of anaphoric relations including both non-referring and referring NPs. 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).

Non-referring. This includes mentions with a specific syntactic or semantic function: predication, expletive, idiom, incomplete or fragmentary expression, quantifier, and coordination. Following ARRAU, we keep all these types of non-referential mentions.

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.[6] This distinction signals whether an entity is mentioned a first or a subsequent time.

Referring mentions can form coreference chains, a group of mentions pointing to the same entity, a central construct in the anaphora 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).

One key principle in these theories is that some referring expressions are used to introduce entities (discourse-new) and some others to refer back to them (discourse-old). In situated dialog, in addition to the textual context, the image provides additional context, constraining the amount of referents and their perceived status by the participants depending on the task in which they are presented (Allopenna et al., 1998). We illustrate this contrast with (2) below. Typically, pronouns are the form of choice for discourse-old entities that have been previously introduced by another expression with lexical meaning. The text in (2), however, starts with *It*. This is possible because the image provides the context and this source of reference ought to be accounted for differently in the annotation than a typical discourse-old case referring back to a *phrase* or *segment* antecedent 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 keep these cases distinct, we introduced the value *task* for the *It* in sentence 1. This means that a discourse-old entity can have distinct types of antecedents: *phrase*, *segment*, or *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...".

3.2.1 Bridging

Another referential relationship included in the ARRAU guidelines is bridging, an associative relationship between two mentions (Versley et al., 2016). When a mention is referential, our annotation indicates whether it is also a related object of some other entity. The *Tell-me-more* corpus is rich in examples of the *part-of* bridging relationship: "An object that stands in a part-of relation to an object previously mentioned" (Artstein and Poesio, 2006). Since the corpus uses pictures of different rooms in a house, after a room is introduced, a series of objects belonging to that room follow, creating many opportunities for using a bridging reference mechanism. For instance, imagine your surprise if the second sentence of example (3) started with *the toaster* instead of *the bed*. Coherence will be immediately broken.

[5] Since the texts are in English, most NPs are marked as neutre.

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

(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.

3.3 Grounding

The ARRAU scheme provides a basic grounding scheme that serves our purposes well (Artstein and Poesio, 2006). In this scheme, the objects in an image have a pre-determined id which can be associated with the text mentions of that object. In our annotation, we take the labels of the bounding boxes as the objects ids. We also differentiate between visible objects with a bounding box and visible objects without a bounding box. For all objects with a corresponding bounding box, the specific object id is linked to its mention in the text.

For bridging references, mentions in a *part-of* relation which do not have a bounding box of their own are grounded to the object that they are a part-of. For example, if the object 'the base of the bathtub' does not have a bounding box, but the object 'the bathtub' does, then 'the base of the bathtub' is grounded to 'the bathtub'.

4 Measuring Agreement

This release contains 500 annotated documents by one annotator and 50 annotated by two. In this section, we detail the computation of the inter-annotator agreement (IAA) using the 50 documents which have been doubled annotated.

Computing IAA for coreference resolution is non-trivial, as annotators need to decide on the mentions boundaries and also which ones belong together in a chain. Following Passonneau (2004), we report Krippendorff's α with weighted δ:

$$\alpha = 1 - \frac{pD_O}{pD_E} = 1 - \frac{rm-1}{m-1} \frac{\sum_i \sum_b \sum_{c>b} n_{b_i} n_{c_i} \delta_{bc}}{\sum_b \sum_c n_b n_c \delta_{bc}} \quad (1)$$

Where m is the number of annotators, and r is the number of coding units, i.e., mentions. For every pair of mentions b and c, δ_{bc} is the distance between the sets formed by their tokens; n_{b_i} is the number of times the value b was assigned by each annotator to each mention i. The distance between the mentions' tokens is 0 when the mentions' tokens are identical, 0.33 when one set subsumes the other, 0.67 when one intersects the other, and 1 when they are disjoint.

To compute Eq. 1, we code our annotations as described in Passonneau (2004), who relies on a predefined number of coding units in order to compute the set distance δ between mentions. Since we do not have predefined mentions because annotators were asked to identify the mentions boundaries by hand, we compute IAA at the token level. This means that our scores are potentially penalized because irrelevant tokens are treated as their own sets.[7]

We compute Krippendorff's α per document and obtained an average of 0.5550. There is a lot of variation, however, with the lowest $alpha$ value at 0 and the maximum at 1, and $\sigma = 0.2263$. Results per document are reported in Table 2.

Doc. id	α	Doc. id	α
8	0.6925	220	0
10	0.4669	237	0.5635
15	0.5641	245	0.6277
26	0.5807	249	0.6212
34	1	251	0
40	0	253	0.6285
53	0.5084	260	0.5921
55	0.7038	266	0.6061
57	0.5955	302	0.6293
62	0.622	311	0.8971
74	0.723	316	0.6737
81	0.6864	340	0.6748
83	0.393	372	0.6146
93	0.6359	387	0.6215
102	0.5319	406	0.1689
107	0	411	0.7609
115	0.4806	416	0.5965
136	0.6077	440	0.5316
163	0.6058	444	0.7366
167	0.6214	445	0.6853
168	0	457	0.4266
176	0.6434	465	0.717
186	0.3105	477	0.7302
196	0.6759	488	0.6225
198	0.7137	500	0.661
average	0.5550		

Table 2: Krippendorff-α for 50 documents double annotated with coreference information following the ARRAU corpus guidelines.

The IAA results obtained are very mixed. The low scores of some documents are partly explained by our choice to do the mention identification completely by hand. This means that the annotators had to decide the boundaries of each mention, yielding

[7] As an illustration, consider the example in (4). Here the tokens *{Mostly, is, is, has, a, and, and, , is, on, is}* are left non-annotated by annotator A; while *{Mostly, is, is, has, and, and, is, on, is}* by annotator B. This is expected as they do not form part of any of their identified mentions, but by scoring at the token level, each set would then be taken as forming a "mention" and hence compared.

imperfect matches even if they agreed on the underlying mention. In the future, we plan to process the text with an automatic mention detection tool and measure our annotations with respect to the tool's output.

4.1 Examples

In this section we present two examples of documents with αs of 0.5807 and 0.

(4) 1. Mostly [[this room]] is [[a bed]]. 2. [[There]] is [[a lamp on a small white nightstand next to the bed.]] 3. [[The bed]] has [a [light blue bed skirt]] and [[white comforter]] and [[4 white pillows]]. 4. [[There]] is [[a blue dresser with a lamp] on [it]]. 5. [[There]] is [[a full length window with vertical shades]].

In this example, we consider the maximal spans for each annotator.[8] Annotator A's annotations are shown with cyan brackets while Annotator B's with blue ones. The example shows that they agree in almost all the boundaries, with disagreements only on sentence 3 *a* and sentence 4 *it*. This also creates a disagreement with the corresponding coreferential chain: for annotator A, the *it* in sentence 4 is coreferential with *a blue dresser with a lamp*; for annotator B, this is part of the singleton *a blue dresser with a lamp on it*.

An α score of 0 occurs when the document does not have any chains, or when at least one of the annotators decided not to annotate anything. This scenario happens when the quality of the text data is unsatisfactory (5).

(5) 1. two beds 2. blue wall 3. three paintings 4. one window 5. tan wall

5 Differences with ARRAU

The annotation guidelines for ARRAU were designed to include a broad range of anaphoric phenomena found in many genres. Our documents are much simpler and the scale of our annotation much smaller, at least at the moment. Issues included in ARRAU but not included here comprise genericity, min words arguments (the head word of a mention), grammatical function, embedded arguments, and any type of complex structure requiring automatic parse of the texts.

[8] Mentions may contain embedded mentions.

6 Conclusion

In this paper, we presented the first release of a portion of the *Tell-me-more* corpus manually annotated with anaphora information according to standard guidelines used for the task of coreference resolution. We also presented IAA scores on 50 documents annotated by two people with training in computational linguistics. Our resource is the first of its kind, although its size is small. However, we believe that it can support linguistic studies about the relationship between textual anaphora and reference to objects, and that it can contribute to research on bridging reference. In addition, it can be used as validation data for automatic methods developed for grounding the entities in the text to the image. This is still work in progress and we look forward to future cycles of revisions and updates of our guidelines in the near-future.

Acknowledgements

The authors thank Sebastiano Gigliobianco for his support setting up the MMAX tool and annotating a large portion of the data. We also thank Philine Huß for her annotation work.

References

Paul D. Allopenna, James S. Magnuson, and Michael K. Tanenhaus. 1998. Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language*, 38(4):419–439.

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).

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.

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.

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.

Prashant Jayannavar, Anjali Narayan-Chen, and Julia Hockenmaier. 2020. Learning to execute instructions in a Minecraft dialogue. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2589–2602, Online. Association for Computational Linguistics.

Mandar Joshi, Omer Levy, Luke Zettlemoyer, and Daniel Weld. 2019. BERT for coreference resolution: Baselines and analysis. 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 5803–5808, Hong Kong, China. Association for Computational Linguistics.

Sopan Khosla, Juntao Yu, Ramesh Manuvinakurike, Vincent Ng, Massimo Poesio, Michael Strube, and Carolyn Rosé. 2021. The CODI-CRAC 2021 shared task on anaphora, bridging, and discourse deixis in dialogue. In *Proceedings of the CODI-CRAC 2021 Shared Task on Anaphora, Bridging, and Discourse Deixis in Dialogue*, pages 1–15, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yuval Kirstain, Ori Ram, and Omer Levy. 2021. Coreference resolution without span representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 14–19, Online. Association for Computational Linguistics.

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)*.

Mingxiao Li and Marie-Francine Moens. 2021. Modeling coreference relations in visual dialog. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3306–3318, Online. Association for Computational Linguistics.

Jiacheng Liu and Julia Hockenmaier. 2019. Phrase grounding by soft-label chain conditional random field. 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 5112–5122, Hong Kong, China. 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.

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

Rebecca J. Passonneau. 2004. Computing reliability for coreference annotation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2017. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *International Journal of Computer Vision*, 123:74 – 93.

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.

Massimo Poesio, Maris Camilleri, Paloma Carretero-Garcia, and Ron Artstein. 2021. Arrau 3 annotation manual.

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.

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).

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

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.

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.

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.