Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 2396–2404
Marseille, 11–16 May 2020

Eye4Ref: A Multimodal Eye Movement Dataset of Referentially Complex Situations

Özge Alaçam, Eugen Ruppert, Amr Rekaby Salama, Tobias Staron, Wolfgang Menzel
University of Hamburg, Department of Informatics
Vogt-Koelln Strasse 30, 22527, Hamburg, Germany
alacam,ruppert,salama,staron,menzel@informatik.uni-hamburg.de

Abstract

Eye4Ref is a rich multimodal dataset of eye-movement recordings collected from referentially complex situated settings where the linguistic utterances and their visual referential world were available to the listener. It consists of not only fixation parameters but also saccadic movement parameters that are time-locked to accompanying German utterances (with English translations). Additionally, it also contains symbolic knowledge — contextual — representations of the images to map the referring expressions onto the objects in corresponding images. Overall, the data was collected from 59 participants in three different experimental setups (86 systematically controlled sentence–image pairs and 2024 eye-movement recordings). Referential complexity was controlled by visual manipulations (e.g. number of objects in the scene, visibility of the target items, etc.), and by linguistic manipulations (e.g. the position of the disambiguating word in a sentence, of the relative clause attachment). This multimodal dataset – in which the three different sources of information, namely eye-tracking, language, and visual environment, are aligned – provides testing possibilities for various research questions not only from a linguistic but also from a computer vision perspective.

Keywords: eye-movement parameters, multimodality, situated language understanding, reference resolution

1. Reference Resolution and Meaning Extraction

Limiting communication to a single mode requires completeness of that mode, and spontaneously spoken words typically lack completeness, since we use language in combination with other modalities to communicate what we want to convey (Tversky, 2014). Therefore, incorporating other cues in relation with the spoken utterance gets us closer to its meaning. A growing body of literature demonstrates that the human language processing system integrates information from various modalities to extract the meaning of the linguistic input accurately, e.g. by resolving references early while the sentence is unfolding and even by re-constructing the meaning from noisy / missing input (MacDonald and Seidenberg, 2006; Altmann and Mirković, 2009; Knoeferle et al., 2005; Tanenhaus et al., 1995).

The facilitating effects of non-linguistic cues like prosodic or contextual cues on the performance of Natural Language Processing (NLP) solutions for reference resolution, disambiguation or meaning recovery have already been well established in the literature (Coco and Keller, 2015; Snedeker and Trueswell, 2003; Ferreira et al., 2013; Spivey and Huette, 2013; Louwerse, 2008), also see Alaçam et al. (2020) for a review. In recent years, with the advancements in eye-tracking technology, incorporating eye movements of a speaker or a listener starts to be seen as another beneficial tool in predicting / resolving which entity in a complex visual environment is referred to in the utterances (Mitev et al., 2018; Koleva et al., 2015).

Uni-modal approaches like language models (low-level N-gram approaches or employing higher level methods like dependency structures) can be quite useful for extracting the intended meaning from text-only material (Asnani et al., 2015; Bickel et al., 2005; Mirowski and Vlachos, 2015; Gubbins and Vlachos, 2013). Semantic classification and clustering methods (e.g. word embeddings and ontologies) can also be used for this purpose. However, when it comes to the description of daily activities, the capability for multimodal integration by employing contextual information can be a very important feature in resolving references and / or performing commands, e.g. for a helper robot that aids people in their daily activities. Linguistic distributions alone could hardly provide enough clues to distinguish the action of *bringing a pan* from *bringing a mug*, which is a crucial difference for helper robots.

Such communication with a helper robot usually happens in structurally rich visual environments like the one in Figure 1, which contains several people, animals, cages, tables, pills, toys, etc., some of them even (partially) occluded from the viewer's perspective. Objects in such an environment can be referred to in quite different, sometimes underspecified manners since some of the relevant information can be easily conveyed by the visual environment. Reconstructing the intended purpose requires more or less complex inferences that rely on the available information about the immediate environment and the world in general. Under these conditions, the optimal interplay between language and visual information, as well as with deictic cues like eye-movements is crucial.

However, the dynamics between these two sources are quite task-specific and heavily under the influence of linguistic and situation-specific constraints. Therefore the investigation in this direction requires a rich, systematically controlled and multimodal data set, which is the main contribution of this paper.

2. The Role of Eye-Movement Parameters on Reference Resolution

As shown by Koleva et al. (2015), listener gaze can be a highly beneficial tool to predict which entity is referred to in the sentence and to understand the intention of the listener when the targets and their referentially possible competitors

are located closely.

A gaze-contingent natural language understanding system may react to changes in its environment by tracking the probability of the fixations per each item in the scene over time. However, as shown by Henderson and Smith (2009), the success of such a system that aims to identify the attended objects is highly dependent on utilizing a combination of several fixation parameters such as fixation location and duration instead of relying on one parameter. Another relevant finding from Koleva et al. (2015) in terms of combining various parameters to enhance the performance of the system indicates that gaze only benefits the model when it is combined with situation-specific features of the current scene.

In this study, we provide a rich set of eye-movement parameters including fixation- and saccade-based parameters. So depending on the research question and domain, a researcher who would like to use the *Eye4Ref* dataset can choose among different eye-movement parameters presented along with the time-locked verbal information and situation-specific information, and even easily calculate / infer some additional features based on the existing ones.

3. Use Case Scenarios

In line with the aim of psycholinguistic experiments where eye-movement parameters were collected (described in the following section), this dataset can be quite useful to train and evaluate computational solutions targeting reference resolution, disambiguation and meaning recovery due to its rich gold-standard annotations. For example, developing a data-driven incremental parser that also successfully incorporates the eye-movements of the speaker for the tasks described above requires a richly annotated dataset that exhibits the multimodal nature of the task. As mentioned in the literature above, combining the eye-movement parameters with linguistic labels and contextual representations to a large extent enhances reference resolution, but the questions of when, or to what extent (in terms of increasing referential complexity) the benefit of cross-modality still stands can be only answered by a systematically manipulated set of experiments.

Another interesting use case scenario of this dataset is a gaze-contingent language understanding system, where the eye-movements of the speaker are incorporated into the parsing solutions in real time. Such a dataset can be used to extract features that represent optimal real-time eye-movement parameters to guide / enhance parsing solutions. Many of the eye-tracking technologies on the market employ a sufficient frequency to enable gaze-contingent applications. In other words, based on a fixation made inside of a pre-defined area-of-interest, triggering an action is possible – e.g. informing the parser in real time about the eye-movement parameters. Then, the parser can use this information to anticipate which entity in the context will be mentioned.

In addition to its contribution for the fields of language understanding and multimodal communication, the object properties provided in this dataset can also be a valuable resource for computer vision, since every object in the images has been annotated with the commonly used properties and annotation guidelines as will be elaborated in the following sections. For example, *color* property, as one of the most basic and frequently used object properties, plays a crucial role in visual saliency detection and hence it has been annotated in a fine-grained way in the dataset.

4. Experimental Setup for Data Collection

Eye4Ref is a rich multimodal dataset of eye-movement recordings collected from three psycholinguistic studies. All experiments address varying referentially complex situated settings where the linguistic utterances are presented with their visual referential world. In this section, we summarize the experimental settings used in all experiments. Experiment-specific information such as linguistic and visual manipulations will be discussed in the upcoming subsections.

Participants In total, 62 students from the University of Hamburg participated in three experiments (*Mean age* = 24.3, *SD* = 5.7, *21 female*). The data from three participants was excluded from the dataset due to insufficient tracking. The experiments were conducted in German, and all participants were native speakers. They all had normal or corrected vision and were paid or given course credit to participate.

Apparatus The stimuli were displayed on an *SR Eyelink 1000 Plus* eye tracker with a sampling rate of 1000 Hz, monocular recording with chin rest apparatus, integrated into a 17-inch monitor with a resolution of 1280×1024 pixels.

Procedure In all experiments, using the visual world paradigm with a simple look-and-listen task, we presented participants with referentially complex images (in total 86 images) with accompanying spoken sentences. The experiments started with filling out the written consent form and demographic data form. Afterwards, instructions were given in a written format followed by familiarization trials. Then, visual calibration with 9 dots was performed. Each trial (presented in a randomized order) began with a drift correction, and later a simple fixation cross located at the middle-bottom of the screen (where there is no overlap between any objects in the following scene) was presented for 3 sec. Next, a visual scene was presented[1] before the onset of the spoken sentence. This preview gives the comprehender time to encode the visual information in advance of the linguistic information being presented. So, visual attention is intended to be free of recognizing the objects of the visual context during language processing. Finally, the spoken sentence was presented accompanying the image. In all experiments, participants were asked to examine the scene carefully and to attend the information given in the audio. A trial ended 4 sec after the offset of the sentence.

Pre-prossessing of the spoken material The sentences in all experiments were recorded by a male native speaker of German at a normal speech rate. Since intonational differences between different linguistic entities have been

[1] For 10 sec in Study-I and II, and 4 sec in Study-III; The viewing time has been set w.r.t. the number of the objects in the scene.

found to have a significant effect on reference resolution (Coco and Keller, 2015; Snedeker and Trueswell, 2003), the intonational breaks that may bias the interpretation have been equalized by using *Audacity* [2].
The average word duration is 624 msec for the content words (*SD* = 236 msec) and 324 msec for the function words (*SD* = 126 msec). The median sentence length is 9 words and the average sentence duration is 6050 msec.

Image Construction and Visual Complexity In order to have flexibility and high control over the systematically manipulated items (in terms of the properties, and spatial arrangements), synthetic images have been created with the *SketchUp Make* Software[3] and all 3D objects were exported from the original SketchUp 3D Warehouse.
The image dimensions are 1250×840 pixels. Moreover, target entities (objects and agents) depending on the task are located in different parts of the visual scene for each stimulus. It should be noted that for developing a computational model of situated language understanding, the symbolic contextual representations of images are sufficient. However, including the images in the dataset has several advantages: First, it makes the list of contextual representations easy to relate with the depicted world for researchers who are going to use this dataset. Second, the images are crucial to conduct comparable experimental studies with human subjects. Third, in order to implement a full pipeline from object detection to reference resolution, an automatic extraction of contextual representations from the images is another crucial and challenging task; and this dataset can provide manually annotated data (gold standards) that could be used to develop and evaluate such a system.
Contextual representations were annotated by two coders. Inter-rater reliability was calculated by Cohen's kappa. The results revealed a value of .71 which indicates a substantial inter-rater agreement. All annotations that were not agreed upon at first were discussed with a third annotator.

4.1. Study-I (higher referential complexity)

This study's focus lies on the role of visual modality in semantic disambiguation of relative clause attachments (RC). Each stimulus is accompanied with a target sentence, a distractor and a connective sentence in between, as illustrated below.

- "It is a mug on a coffee table that she damages (*the target sentence*). Then, the man reaches a decision (*the connector sentence*) . He cleans the pillow on the armchair (*the distractor sentence*)".

In this study, all target sentences are syntactically ambiguous, while only half of them carry semantical ambiguity as well, as exemplified in [1] below.
Syntactical ambiguity is introduced by the German language which has three grammatical genders. In German, each noun is either feminine (*f*), masculine (*m*), or neuter (*n*). In a sentence that contains a relative clause attachment, the gender of the relative pronoun has to be the same as the gender of its antecedent. To illustrate, the relative clause

[2]http://www.audacityteam.org/
[3]http://www.sketchup.com/

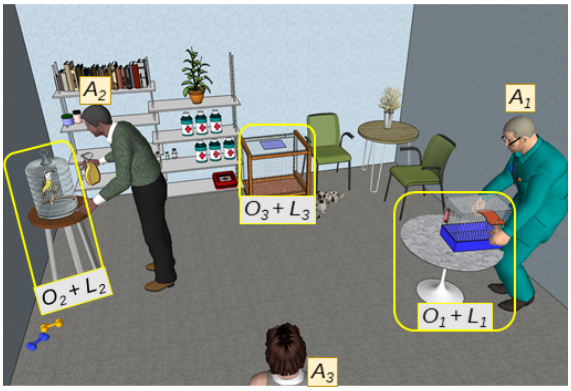

Figure 1: The image for the sentence 1C. Referential complexity: *high* (A_1-O_1-L_1: Target entities; A_2-O_2-L_2: Congruent distractors; and O_3-L_3: Incongruent distractors)

in a neuter form *das* in the sentence [1A] is in a syntactic agreement with both preceding nouns; ein Buch (*a book*) and ein Sofa (*a sofa*).
In semantically unambiguous cases, the verb is semantically congruent with either NP (*high-attachment*) or PP (*low-attachment*), as exemplified in sentence [1A]; among two options, the word *book* would have high preference rate for the action *read*. In the semantically ambiguous cases, both can be licensed by the verb, as in sentence [1C]: *he* can move the *cage* or the *table*. In such cases, correct reference resolution cannot be achieved based on linguistic information alone, hence the contribution of visual information is indispensable, see Figure 1.

1A. **Semantically Unambiguous and High-attached.**
Da befindet sich ein Buch*(n)* auf einem Sofa*(n)*, das*(n)* er ruhig liest.
(*It is a book on a sofa that he reads quietly*)

1B. **Semantically Unambiguous and Low-attached.**
Da ist ein Umschlag*(m)* auf einem Schreibtisch*(m)*, den*(m)* er langsam zusammenbaut.
(*It is envelope on the desk that he slowly assembles.*)

1C. **Semantically Ambiguous and High-attached.**
Da befindet sich ein Käfig*(m)* auf einem Tisch*(m)*, den*(m)* er achtsam trägt.
(*It is a cage on a small table that he moves carefully.*)

1D. **Semantically Ambiguous and Low-attached.**
Da ist eine Flasche*(f)* in einer Tasche*(f)*, die*(f)* sie schnell öffnet.
(*It is a bottle in a bag that she opens quickly.*)

– To fully comprehend the described event, as of the RC-verb onset, the reference resolution is determined by backward (off-line) processing of the information which has been already uttered in interaction with visual information.

– One should note that in all cases, the visual scene depicts only the correct interpretation without an ambiguity.

– To increase the referential computation, each referenced item is provided with its distractor competitors; three

agents, three possible referent objects, three locative objects and other background objects. Perceptual saliency of the target and the distractors are kept similar.

Five sentences for each category were created, resulting in 20 scene–sentence pairs and 300 eye-movement recordings (15 participants $\times$ 20 scenes).
In addition to the systematically controlled sentence-image parts, the multimodal data for the distractor sentences, where the referential complexity of the target objects in the scene is not manipulated, are also provided in the Eye4-Ref dataset.

4.2. Study-II (moderate referential complexity)

Unlike the previous study, the images do not contain any animate characters in this study. Thus, only the spatial relations and object properties exist to resolve which object in the sentence is being referred to.
The visual complexity is manipulated by the presence of occlusion. The target objects are visually presented as fully-visible or partially-occluded, as can be seen in Figure 2. Moreover, the position of the identifying modifier in a sentence is set as either before the NP [2A] or after the first instances of the NP [2B]. By this manipulation, either the *type* information of the object or visual properties like *color*, *shape*, *material* or *size* have a key role in disambiguation.

2A. Bring mir *den blauen* Becher vom Tresen.
Bring me the blue mug from the counter.

2B. Bring mir den Becher, *den blauen* Becher, vom Tresen.
Bring me the mug, the blue mug, from the counter.

26 subjects participated in this study (in a between-subject design). In total, 32 scenes were accompanied with 2 linguistic variations of the same sentence (64 sentence recordings). As shown in Figure 2, each scene has a target object with a specific attribute value (e.g. *blue* as color) and two other distractor objects of the same kind with different values of the same attribute (e.g. one *purple* and one *yellow* mug). Moreover, some other incongruent distractor objects (e.g. three vases of all colors) have been also included to increase the referential complexity.

- The sentence can be fully disambiguated after the NP. PP only has a complementary role, in case the disambiguation is not completed after NP due to visual complexity.
- The sentences in both conditions have the same amount of referential selections, only the sentence structure is slightly different. In other words, the first occurrence of the word *blue* or *mug* narrows down the search space to three blue items (represented by the *color* attribute) or three mugs (by *type*) respectively.
- Occlusion has been set to hinder the *type* information (i.e. the handle of the mug is occluded) but not the *color* property. In the occluded setting, [2A] is expected to cause early reference resolution compared to [2B].

Figure 2: The image for the sentence 2A and 2B. While the full image displays the fully visible target, the image part at the bottom-left corner illustrates the partially occluded target object. Referential complexity: *Moderate*

4.3. Study-III (low referential complexity)

The stimuli in this study are designed as a control group for a psycholinguistic study that focuses on human preferences for the reconstruction of acoustically unclear sentence parts (in German), more specifically on obtaining prior probabilities of three types of grammatical words: two common prepositions of location (*on* and *next to*) and the negation particle (*not*), given certain visual arrangements of the objects (Alaçam, 2019). Therefore, as in the original study, a constant background noise (a sound recording from a restaurant) accompanies the entire spoken sentence.
The reason for producing sentences with a repair phrase, which conveys additional information, lies in having a referential competition in the environment. Otherwise, the production of such sentences would yield unnecessary overspecified utterances. Nevertheless, the scenes in this study have been designed to have relatively low referential competition as compared to previous studies (i.e. there are two tables in the scene, making the decision a binary task instead of a multinomial one).
All sentences have the same structure except the negation / preposition part as given below. The sentences start with a verb in an imperative form preceding an object (*NP*) and a prepositional phrase that specifies the goal location (*PP*). Then, the sentence continues with a disfluency (*umm*) and a repair / complement part consisting of a negation or one of the two prepositions of location.

3A. Stell den Becher auf den Tisch, umm [nicht] den blauen. (both low and far attachment)
Put the mug on the counter, umm [not] the blue one.

3B. Stell den Becher auf den Tisch, umm [auf] den blauen. (only low attachment)
Put the mug on the counter, umm [on] the blue one.

3C. Stell den Becher auf den Tisch, umm [neben] den blauen. (only far attachment)
Put the mug on the counter, umm [next to] the blue one.

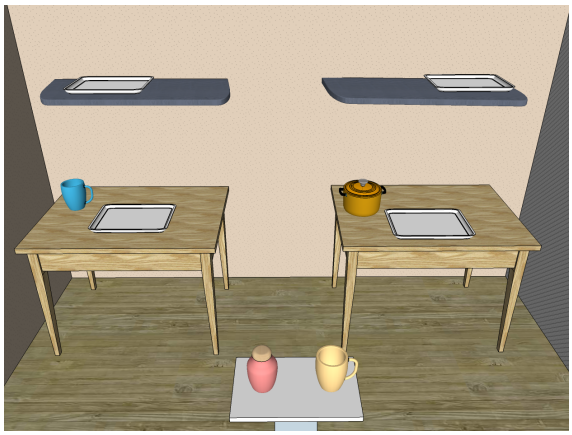

Figure 3: The image for the sentence 3A-B and C, Referential complexity: *low*

In this setting, the repair / complement may have three different syntactic roles; referring back to the OBJECT which is the mug (with *not*), referring back to the ADVERBIAL which is the table (with both *on* and *not*) or providing new complementary ADVERBIAL which is another mug (with *next to*). Due to filling different roles, all possible linguistic interpretations require different parsing results. In all cases, the object referred to in the repair/complement part shares either the property (e.g. blue) or the object class (e.g. mug) with the target object or location.

In this study, one third of the sentences involve a negated statement. To our knowledge, there is no comparable dataset that involves *negated statements* presented in a systematically manipulated situated language setting. Identifying the scope and focus of negation is one of the challenging issues in the NLP community (e.g. *SEM 2012 shared task, Morante and Blanco (2012)). The psycholinguistic literature agrees that sentences containing negation are harder to interpret than affirmative sentences (Orenes et al., 2014; Khemlani et al., 2012; Kaup et al., 2006; Lüdtke and Kaup, 2006; Carpenter and Just, 1975). On the other hand, it has been shown that when negation is supported by the right contextual support, the positive argument no longer needs to be represented, yielding faster verification compared to no-context situations (Tian et al., 2016; Dale and Duran, 2011; Nieuwland and Kuperberg, 2008).

Although many more different visual arrangements are possible, for the sake of systematicity, the visual conditions (the object properties and their spatial relations among each other) are limited to five scene arrangements in this set. Figure 3 illustrates the first condition, that conveys all possible interpretations for all the focus-of-interest fillers. Details of the experimental stimuli can be found in Alaçam (2019) and also in the repository description. The number of objects in the scenes is limited to eight and one additional object is used in one of the scene variations. For each visual condition, six different visual scenes were designed, resulting in 30 trial scenes.

5. Dataset

With this dataset, we aim to provide researchers from various domains with a richly annotated multimodal data set. Our multimodal dataset consists of highly intertwined information coming from three different sources: language, eye-movements and scenes. In this section, we explain the design of the dataset.

5.1. Data Preprocessing

The dataset is made available in two formats; (i) *fixation-based format* where each row in the dataset belongs to one individual fixation, and (ii) *time-series format* where each row contains the location of fixation in 20 msec bins. The *time-series* format enables a more fine-grained alignment to the unfolding sentence, therefore it is more suitable to investigate how reference resolution develops in time while the sentence unfolds. In the *time-series* format, the fixations were coded as binomial w.r.t. whether the object is fixated or not. The average values for the bins were directly calculated by *SR Eyelink Data Viewer 4.1.1 Software*.

The trials are further divided into two main phases; the first one corresponds to the free-view phase before the sentence onset, and the second one is the task specific period after the onset of the sentence until the end of the trial. The time window is shifted forward 200 ms in order to account for the time required to initiate eye movement (Matin et al., 1993). The dataset contains 144 unique lexical items (98 nouns, 26 adjectives and 20 verbs) with varying word length and 578 objects that belong to 80 object categories. In total, it consists of fixation distributions of 2024 eye-movement recordings (see Table 1). The time series that contain more than 35 % empty values (due to blinks, misses or fixations outside of the recording area) were discarded, resulting in 1867 trials overall.

5.2. Contextual Representations

Cross-modal integration between language and vision can be addressed by various methods. One of the common methods is to relate uni-modal features from different modalities on a conceptual level by using common representations such as semantic role labels (McCrae, 2009; Mayberry et al., 2009; Salama and Menzel, 2018). Semantic roles are linguistic abstractions to distinguish and classify the different functions of the action in an utterance, in other words they are a useful tool to specify *"who did what to whom"*.

We use semantic roles as common representations to establish a relation between semantic and syntactic levels as well as the non-linguistic information. The information carried in the visual modality is represented in a symbolic knowledge base that contains the relationships between objects, characters and actions in the scene. This information has been manually annotated with a triplet notation as `<argument, relation type, predicate>`.

Despite slight variations, this notation can be considered as a common methodology for knowledge representations in various domains: computer vision, robotics, language technologies. With respect to this notation, the `relation type` is one from a predefined set of accepted relations, such as AGENT or LOCATION, while Predicate and Argument are tokens of the input sentence. Currently, we specify the context relations under two categories; event relations (namely AGENT, THEME, INSTRUMENT), and state

	Studies	# of Participants	Sentences	# of Scenes	# of Trials
Ia	Study-I (higher referential complexity)	12	20	20	240
Ib	Study-I (uncontrolled referential complexity)	12*	20	20*	240
II	Study-II (medium referential complexity)	27	32	32	864
III	Study-III (low referential complexity)	20	34	34	680
	TOTAL	59	106	86	2024

Table 1: The number of participants, of scene-scenario pairs and of the trials in total for each study (*: same participants and the scenes as in the previous condition).

relations such as LOCATION, PART-OF, or object properties like TYPE, COLOR, MATERIAL, etc. (e.g., a yellow chewing stick, a paper box, etc.). Table 2 shows one exemplary semantic annotation for the visual scene displayed in Figure 1. There, *"Man$_1$"* is the AGENT, who performs the carrying action, *"Cage$_1$"* is the THEME, the entity undergoing a change of state, caused by the action.

While labeling the sentences, semantic roles such as GOAL and SOURCE are particularly useful to carry the direction of the action. On the other hand, the contextual representations for a static image only allows the use of LOCATION. Still, those roles are crucial for annotation of dynamic scenes, which are currently not the part of the dataset.

The contextual representations provided in this dataset include:

- Object ID (*Study#,Image#,Entity#*). Every object in the dataset has a unique object ID to establish the relation between the contextual representations and the eye-movement recordings.
- Context-specific entity ID. e.g. *mug$_1$*
- Location Relations (*on, next to, above, below, inside*)
- Object Properties (*type, color, size, shape, part-of, texture, visibility*)
- AOI-Pixel size: Boundary of the objects (*Area-of-Interest*) is marked and defined with *SR Data Viewer Software*.

There are some cases where defining AOI for each individual item is not reasonable, for example in case there are many books located on a shelf side by side. Then, only one AOI is defined for all the books, and the *count* relation is set to *many* to represent its plurality.

For illustration, the target object which has been mentioned in Sentence [2A and 2B] and depicted in Figure 2 has two different ID types: a unique object ID ($S001I002E001$) and a context-specific entity ID (mug_1) displaying following contextual representation:

- mug as *type*
- blue as primary color *colorP*, 190 as *Hue*, 46 as *Saturation*, 82 as *Value*
- *loc$_{ON}$* relation with counter$_1$
- *loc$_{Next_To}$* relation with icebucket$_1$
- fully visible as *visibility*
- ...

Among object properties, special attention has been given to the *color* property since it could be one of the common and basic properties for reference resolution, also for the field of computer vision, for example to detect bottom-up visual saliency. Therefore, the color attribute of each object has been labelled with two parameters; (i) textual color labels (e.g. blue, dark gray etc.) and (ii) HSV values[4] (hue, saturation, value), which are well suited to represent color information for data-driven computational solutions. On the other hand, in order to account for the objects that display more than one color, three more color-related properties are utilized; a primary color (*colorP*), secondary color *colorS* (if any, same sub-parameters as above), and *texture* (yes or no) are specified.

It should be noted that the relative properties like for example a *size* attribute require an additional step to be able to appropriately carry this relation. In triplet notation, where the relation is a limited set of functions, to represent relative properties that require another object to be compared is a tricky issue. To illustrate, among the following representations $(plant_1, size, big)$, $(plant_2, size, small)$, $(candle_1, size, big)$, $(candle_2, size, big)$, the comparison should be done among the instances of the same object category (namely among the items that have the same *type* information); *candle_1* is bigger than *candle_2*.

Table 2 contains the relations for the most complex setting among the three studies with some characters and background objects and interaction among them (see Figure 1). As visual complexity increases, the number of the contextual representations for the scene increases as well. It should be noted that this relation space can grow further depending on the granularity level of the relations that have been chosen for annotations.

To wrap-up, the current version of our multimodal dataset in German that we constructed with the aim of studying disambiguation and structural prediction from both psycholinguistics and computational linguistics perspectives contains the following items for each scenario in the dataset:[5]

- meta-data information for each trial
- sentences in German (with English translation)
- gold standard linguistic annotations (POS, lemma, dependency structures)
- images which the sentences refer to

[4] https://psychology.wikia.org/wiki/HSL_and_HSV

[5] The dataset can be accessed from https://gitlab.com/alacam/eye4ref

Event Relations	*State Relations*
1. *(Man_1, AGENT, carry_1)	1. (Bird_1, LOC_{IN}, Cage_2)
2. (Man_2, AGENT, spray_1)	2. *(Table_1, $\text{LOC}_{\text{NEXTTO}}$, Man_1)
3. (Woman_1, AGENT, observe_1)	3. (Table_2, $\text{LOC}_{\text{NEXTTO}}$, Man_2)
4. (Bird_1, THEME, spray_1)	4. *(Cage_1, LOC_{ON}, Table_1)
5. *(Cage_1, THEME, carry_1)	5. (Cage_2, LOC_{ON}, Table_2)
6. (Cage_2, THEME, verköstigen_1)	6. *(Uniform_1, PART-OF, Man_1)
7. (Spray_1, INSTRUMENT, spray_1)	7. (Uniform_1, COLORP, green)[6]
	8. (Cage_1, SHAPE, rectangular)
	9. (Cage_2, SHAPE, round)
	10. ...

Table 2: A partial list of the contextual representations in a triplet notation `<argument, relation, predicate>` for the scene illustrated in Figure 1 (*: the relevant relations for the sample sentence [1C]).

- contextual representations of the images
- an audio file and a data file with marked onset/offsets (in msec) of each linguistic entity in the sentence
- eye-movement parameters

The eye-movement recordings for each study enhanced with linguistic labels and contextual information are provided in separate `CSV` files.
To summarize, the target and all the other distractor objects in the images are specified by using common-sense notation and object properties (Dale and Reiter, 1995). Universal POS-tags and CoNLL format are used to label the linguistic modality. The eye-movement parameters are exported from the *SR Eyelink Data Viewer* software and the normalization calculations are explained in the description files along with the dataset.

5.3. Meta-Data Parameters

- *Study ID.* Each study has its unique ID.
- *Condition ID.* Each study has varying number of conditions depending on its design.
- *Audio file name.*
- *Image file name.*
- *AOI file name.* Area-of-interest file that contains all information regarding the AOIs for the respective image.
- *Participant ID.* Each participant has a unique ID.
- *Participant Age.*

5.4. Linguistic Parameters

The words in the spoken sentences were tagged by using Universal part-of-speech tags. The lemma form of the words are provided both in German and English. However, in German, depending on the gender of the noun in the preposition phrase, the preposition and the article can be represented as a fusion word. For example, the fusion word *vom* is commonly used instead of its separate form *von dem* ("from the" in English). Such preposition-article combinations in the sentences are marked with ADP tag (*adposition*) as suggested in the Universal Tagset Conversion Table [7]. Moreover, the dependency structures are represented in the CoNLL-U Format.

- Lemma form (*becher*)
- Lemma form in English (*mug*)
- Universal POS tags (*NN*)
- Dependency structures (only for immediately connected nodes (head-child relationships))
- Normalized Parameters
  - *The length of the linguistic entity.* from the word onset to its offset
  - *Time passed until the current entity.* e.g. from the sentence onset to the onset of the current entity
  - *The length of the sentence.* e.g. from the sentence onset to its offset

5.5. Eye-Movement Parameters

As elaborated in Section 5.1, the eye-movements are provided in two different formats: *fixation-based* and *time-series* format. Since eye-movements are quite individual, relying only on parameters like the number of fixations, average fixation duration, or a peak saccade velocity could be misleading. Therefore these values have been normalized within each trial (for each participant). The description for each of the parameters, which are directly exported from *SR Eyelink Data Viewer*, and of the normalized parameters are explained in the *List of Parameters* documents provided in the repository.
The fixation duration threshold is set to 80 msec, so any fixation shorter than this threshold is excluded. Moreover, the fixations beyond the display area and that occur just after the blinks have also been removed. No merging has been utilized for contiguous short neighboring fixations since this procedure may hinder important aspects of data that could be useful for on-the-fly calculations for gaze contingent systems.

[6] It also contains three more color properties for hue, saturation and value of the color (HSV color space).

[7] https://universaldependencies.org/tagset-conversion/de-conll2009-uposf.html

5.5.1. Fixation-based Parameters

- *Fixation Index* (within-trial parameter)
- *Duration*
- *Start and End time of the fixation* (within-trial parameter)
- *AOIs which have been fixated*
- *Distance to the Nearest AOI*
- *X and Y Coordinates of the fixation*
- *Pupil size*

In real-time gaze-based applications, only the eye-movements made until that point of time are usable to make an inference about the current or the upcoming reference. However, making the calculation based on all the previous eye-movements[8] may not be efficient in terms of computational resources and also does not guarantee accuracy. Therefore, in order to provide a normalized internal index, each time-bin in *the time-series* format was indexed within the fixation that they originally belong to.

- *Normalized Fixation Index*
- *Normalized Fixation Duration*

5.5.2. Saccadic Movement Parameters

- *Saccade Index* (within-trial parameter)
- *Direction* (up, down, right, left)
- *Duration*
- *Amplitude*
- *Average Velocity*
- *Average Peak Velocity*
- *Start and End times of the saccade*
- *X and Y Coordinates of the saccade*
- *Normalized Duration* (in only *time-series* format)
- *Normalized Amplitude* (in only *time-series* format)
- *Normalized Average Velocity* (in only *time-series* format)

5.6. Task-Specific Parameters

- *View Type.* Task-specific role of the object being fixated. For example, the *mug_1* is the target object for the scenario depicted in Figure 2, the *mug_2* is the congruent distractor, and the *vase_1* is the incongruent one.
- *Referential Competition Score.* The number of the shared properties with the target object for the task at hand. As listed in Table 3, e.g., while all the mugs in Figure 2 share the same TYPE information, they differ in COLOR, $\text{LOC}_{\text{NEXTTO}}$, $\text{LOC}_{\text{BELOW}}$ and LOC_{ON} attributes. However, it should be noted that, their difference e.g. on the LOC_{ON} relation are at the entity ID level (*counter_1* vs. *counter_2*), yet they still share the same object type for this relation (*counter*).

[8]The eye-movements made until the current point in time, starting from the beginning of the utterance.

	mug_1	**mug_2**	**mug_3**
	target	distractor	distractor
TYPE	mug	mug	mug
COLORP	blue	purple	yellow
LOC_{ON}	counter_1	shelf_1	counter_2
$\text{LOC}_{\text{NEXTTO}}$	icebucket_1	-	vase_1
$\text{LOC}_{\text{NEXTTO}}$	-	-	faucet_1
$\text{LOC}_{\text{NEXTTO}}$	-	-	teller_1
$\text{LOC}_{\text{BELOW}}$	-	shelf_2	-
$\text{LOC}_{\text{BELOW}}$	shelf_3	-	-

Table 3: Object Properties and Relations for all three instances of *Becher (mug)* displayed in Figure 2.

6. Conclusion

In this paper, we have presented the *Eye4Ref* fully annotated multimodal dataset collected from three different eye-tracking studies on reference resolution and disambiguation tasks in situated settings. The dataset consists of three main components: linguistic labels, eye-movement parameters and the contextual representations of the images. Combined with the images and the audio files, it allows for an end-to-end implementation featuring object detection, speech recognition and NLP tasks like reference resolution, disambiguation and meaning extraction. All components were coded by following commonly used annotation methods. This enables straightforward extension of the dataset and sustains its re-usability by other researchers in the communities. We believe that the dataset will be a valuable resource in the investigation of various research questions of cross-modal language–vision interaction.

7. Acknowledgments

This research was partially funded by the German Research Foundation – DFG Transregio SFB 169: Cross-Modal Learning. We also thank our anonymous reviewers whose detailed comments helped improve and clarify the paper.

8. Bibliographical References

Alaçam, Ö., Xingshan, L., Menzel, W., and Staron, T. (2020). Crossmodal language comprehension – psycholinguistic insights and computational approaches. *Frontiers in Neurorobotics*, 4.

Alaçam, Ö. (2019). Enhancing natural language understanding through cross-modal interaction: Meaning recovery from acoustically noisy speech. In Mareike Hartmann et al., editors, *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 272–280, Turku, Finland. Linköping University Electronic Press.

Altmann, G. T. M. and Mirković, J. (2009). Incrementality and prediction in human sentence processing. *Cognitive Science*, 33(4):583–609.

Asnani, K., Vaz, D., PrabhuDesai, T., Borgikar, S., Bisht, M., Bhosale, S., and Balaji, N. (2015). Sentence completion using text prediction systems. In *Proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2014*, pages 397–404. Springer.

Bickel, S., Haider, P., and Scheffer, T. (2005). Learning to complete sentences. In *European Conference on Machine Learning*, pages 497–504. Springer.

Carpenter, P. A. and Just, M. A. (1975). Sentence comprehension: a psycholinguistic processing model of verification. *Psychological review*, 82(1):45.

Coco, M. I. and Keller, F. (2015). The interaction of visual and linguistic saliency during syntactic ambiguity resolution. *The Quarterly Journal of Experimental Psychology*, 68(1):46–74.

Dale, R. and Duran, N. D. (2011). The cognitive dynamics of negated sentence verification. *Cognitive science*, 35(5):983–996.

Dale, R. and Reiter, E. (1995). Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive science*, 19(2):233–263.

Ferreira, F., Foucart, A., and Engelhardt, P. E. (2013). Language processing in the visual world: Effects of preview, visual complexity, and prediction. *Journal of Memory and Language*, 69(3):165–182.

Gubbins, J. and Vlachos, A. (2013). Dependency language models for sentence completion. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1405–1410.

Henderson, J. M. and Smith, T. J. (2009). How are eye fixation durations controlled during scene viewing? further evidence from a scene onset delay paradigm. *Visual Cognition*, 17(6-7):1055–1082.

Kaup, B., Lüdtke, J., and Zwaan, R. A. (2006). Processing negated sentences with contradictory predicates: Is a door that is not open mentally closed? *Journal of Pragmatics*, 38(7):1033–1050.

Khemlani, S., Orenes, I., and Johnson-Laird, P. N. (2012). Negation: A theory of its meaning, representation, and use. *Journal of Cognitive Psychology*, 24(5):541–559.

Knoeferle, P., Crocker, M. W., Scheepers, C., and Pickering, M. J. (2005). The influence of the immediate visual context on incremental thematic role-assignment: Evidence from eye-movements in depicted events. *Cognition*, 95(1):95–127.

Koleva, N., Villalba, M., Staudte, M., and Koller, A. (2015). The impact of listener gaze on predicting reference resolution. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 812–817.

Louwerse, M. M. (2008). Embodied relations are encoded in language. *Psychonomic Bulletin & Review*, 15:838–844.

Lüdtke, J. and Kaup, B. (2006). Context effects when reading negative and affirmative sentences. In *Proceedings of the 28th annual conference of the cognitive science society (CogSci)*, volume 27, pages 1735–1740. Lawrence Erlbaum Associates Mahwah, NJ.

MacDonald, M. C. and Seidenberg, M. S. (2006). Constraint satisfaction accounts of lexical and sentence comprehension. In *Handbook of psycholinguistics*, pages 581–611. Elsevier.

Matin, E., Shao, K. C., and Boff, K. R. (1993). Saccadic overhead: Information-processing time with and without saccades. *Perception & psychophysics*, 53(4):372–380.

Mayberry, M. R., Crocker, M. W., and Knoeferle, P. (2009). Learning to attend: A connectionist model of situated language comprehension. *Cognitive Science*, 33(3):449–496.

McCrae, P. (2009). A model for the cross-modal influence of visual context upon language procesing. In *Proceedings of the International Conference RANLP-2009*, pages 230–235, Borovets, Bulgaria.

Mirowski, P. and Vlachos, A. (2015). Dependency recurrent neural language models for sentence completion. *arXiv preprint arXiv:1507.01193*.

Mitev, N., Renner, P., Pfeiffer, T., and Staudte, M. (2018). Using listener gaze to refer in installments benefits understanding. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society (CogSci)*, Madison, Wisconsin.

Morante, R. and Blanco, E. (2012). *SEM 2012 shared task: Resolving the scope and focus of negation. In * *SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, volume 1, pages 265–274.

Nieuwland, M. S. and Kuperberg, G. R. (2008). When the truth is not too hard to handle: An event-related potential study on the pragmatics of negation. *Psychological Science*, 19(12):1213–1218.

Orenes, I., Beltrán, D., and Santamaría, C. (2014). How negation is understood: Evidence from the visual world paradigm. *Journal of Memory and Language*, 74:36–45.

Salama, A. R. and Menzel, W. (2018). Learning context-integration in a dependency parser for natural language. In Khaled Shaalan, et al., editors, *Intelligent Natural Language Processing: Trends and Applications*, pages 545–569. Springer International Publishing.

Snedeker, J. and Trueswell, J. (2003). Using prosody to avoid ambiguity: Effects of speaker awareness and referential context. *Journal of Memory and language*, 48(1):103–130.

Spivey, M. J. and Huette, S. (2013). Toward a situated view of language. In Pia Knoeferle, et al., editors, *Visually Situated Language Comprehension*, volume 93 of *Advances in Consciousness Research*, chapter 1, pages 1–52. John Benjamins Publishing Company, Amsterdam/Philadelphia.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., and Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268(5217):1632–1634.

Tian, Y., Ferguson, H., and Breheny, R. (2016). Processing negation without context–why and when we represent the positive argument. *Language, Cognition and Neuroscience*, 31(5):683–698.

Tversky, B. (2014). Visualizing thought. In *Handbook of human centric visualization*, pages 3–40. Springer.