2010-11-01

Visual Salience and Reference Resolution in Situated Dialogues: A Corpus-based Evaluation.

Niels Schutte
Dublin Institute of Technology, niels.schutte@student.dit.ie

John Kelleher
Dublin Institute of Technology, john.d.kelleher@dit.ie

Brian Mac Namee
Dublin Institute of Technology, brian.macnamee@dit.ie

Visual Salience and Reference Resolution in Situated Dialogues: A Corpus-Based Evaluation

Niels Schütte and **John Kelleher** and **Brian Mac Namee**

Dublin Institute of Technology
AI Group

niels.schutte@student.dit.ie
john.d.kelleher@dit.ie
brian.macnamee@dit.ie

Abstract

Dialogues between humans and robots are necessarily situated. Exophoric references to objects in the shared visual context are very frequent in situated dialogues, for example when a human is verbally guiding a tele-operated mobile robot. We present an approach to automatically resolving exophoric referring expressions in a situated dialogue based on the visual salience of possible referents. We evaluate the effectiveness of this approach and a range of different salience metrics using data from the SCARE corpus which we have augmented with visual information. The results of our evaluation show that our computationally lightweight approach is successful, and so promising for use in human-robot dialogue systems.

Introduction

An exophoric referring expression is a referring expression that denotes an object that has not previously been introduced into the linguistic context but that is in the spatio-temporal context of the discourse. This may for example be the physical or visual context. Exophoric referring expressions are particularly important for robot dialogue systems as both the robot and the user may make reference to objects in the visual context.

Dialogue systems in which the system and user share a visual context are especially common in human-robot interaction scenarios. For example, a semi-autonomous tele-operated mobile robot that is controlled through speech by a user who perceives the robot's environment through a camera mounted on the robot will need to understand the user's references to objects seen in the camera view of the robot's environment. We posit that in this scenario the salience of objects plays a particular role. Furthermore, we posit that in a scenario like this, a simple, attention-based salience approach may produce good results.

To evaluate these claims, we use data from the SCARE corpus (Stoia et al. 2008) which features a collection of situated human-human dialogs in which one user directs the actions of a second user in a virtual environment on which both users share a first-person perspective view. We extract all of the exophoric references from these situated human-human dialogues, build a visual model to correspond with the users' perspective at the time of the reference, and resolve the references against this model. We then evaluate the results of our resolution process against gold standard annotations.

The remainder of this paper proceeds as follows. First, related work is discussed. Following this we introduce the corpus used and the additional data generated from it. We then describe the evaluation method and the salience metrics used before presenting and discussing the results of the evaluation. Finally, we draw conclusions on the implications of these results for human-robot dialogues and outline plans for future work.

Related Work

To interpret exophoric referring expressions, a dialogue system needs to have access to a representation of the visual context. If the representation is at a sufficiently abstract level, the process of identifying referents reduces to picking the most likely candidate object from a set of candidates.

In resolving purely linguistic anaphora, some notion of salience is often employed. In Centering Theory (Grosz, Weinstein, and Joshi 1995), for example, the salience of concepts occurring in a sentence are ranked according to their syntactical role, under the assumption that certain roles imply a higher salience than other roles, making them more likely to be the intended referent of a referring expression. In visual domains further information can be used and (Kievit et al. 2001) describe using the visibility of objects to rank the salience of referents. The notion of the salience of an object, however, depends not only on the the properties of the object itself but also on the attentional state of the perceiver. (Kelleher 2006) presents a model where the salience of an object is based on the closeness of the object to the point of visual focus. Similarly, spatial expressions such as prepositions can also be employed to specify objects in a visual scene. In (Gorniak and Roy 2004) an approach is developed to identify objects on the basis of grounding spatial expressions with data from perception. It is worth mentioning that there exist more sophisticated and complex approaches to modeling visual attention than those mentioned here with regard to reference resolution (see (Itti and Koch 2001) for a good overview). However these computationally very expensive processes are not suitable for real-time human-machine dialogue systems.

Approaches such as (Gorniak and Roy 2004) work with a static view of a complete scene. This ignores issues that can arise when the the perceiver navigates inside the environment and perceives only part of the scene from a first person perspective. For example, objects appear or disappear from the perspective depending on the movements of the perceiver and the perceiver can control the perspective by moving to a different position, avoiding perspectives that require complex referring expressions.

In this work we address these issues by basing our evaluation on a dynamic virtual environment in which these sorts of complications arise. We also adopt a strongly simplified approach to suit the real-time human-robot interaction domain: simple metrics for salience are used and simple rules guide attention. In addition simple salience metrics that do not use complex features that are specific to the visual domain may be more easily be applied to non visual sensor data such as output from laser scanners or ultra sound sensors. In the next section we present the data used in our evaluation.

Data

The original data for our experiment is taken from the SCARE corpus (Stoia et al. 2008). The SCARE corpus contains data collected in an experiment focusing on task-based situated dialogue. In the experiment two participants cooperate to fulfill a navigation task in a spatial environment simulated by a game engine. In total 15 recorded dialogs with a total length of about 220 minutes were available. We have augmented this dialogue corpus with data generated from the visual views of the environment which are made available to accompany the dialogues. This section will first describe the SCARE corpus before explaining our augmentations.

In the rest of the text, the first participant in the SCARE experiments will be referred to as "direction follower" (DF), the second participant as "direction giver" (DG). The term "player" will be used to denote the virtual entity that represents the DF in the virtual environment, and the term "game" will be used to refer to each run of the experiment.

The DF was given the task of navigating the environment. The DF perceived the environment from a first person perspective and used a computer keyboard to move around. The DF was given no information about the layout of the world or the details of the task. Instead, this information was given to the DG whose task it was to instruct the DF. The DG was given access to a live feed of the perspective of the DF. Thus, the participants had a shared perspective on the environment. Figure 1 shows a screenshot from a video from the corpus that shows the perspective of the participants on the environment. The participants were allowed to communicate freely through a voice connection.

The virtual environment consisted of a number of rooms that contain cabinets and buttons. The rooms were connected through doors that automatically open if the player approaches. Cabinets could be opened and closed by activating buttons. Buttons were activated by the player walking into them. Some cabinets contained items. To successfully fulfill the task, the participants had to retrieve certain items and move them to different cabinets.

Figure 1: A screenshot from a video recording from the SCARE corpus. It shows the perspective of the DF which is shared with the DG.

Objects were specifically designed so that all objects of a class look the same. This was done to encourage users to use spatial relations in referring expressions instead of simple attributes such as colour or size (Stoia et al. 2008).

The corpus comprises audio recordings and time aligned transcriptions of the dialogues as well as video recordings of the screen of the navigating participant. The transcriptions were annotated for references to objects in the environment. In addition, demo files, that is files that record a specification of all events in a game, were provided for each game. These can be replayed inside the game engine to recreate the original game.

Creating the visual context

The resolution of an exophoric expression requires access to some sort of model of the visual context. However, the SCARE corpus does not directly contain this information, but we can use the game engine and the information about the geometry of the world to recreate it. For this purpose we developed a ray-casting based visibility test that was able to record which objects were visible on the screen at a given time. Ray-casting is a method used in 3D graphics to determine visibility (Foley et al. 1996). It works by sending vectors, or rays, from the virtual eye-point of the observer into the depth of the scene and recording which objects the vectors intersect with. The object closest to the eye-point that a vector intersects with is the object that is visible to the observer along that vector.

To determine which objects are visible in a view of the virtual environment, we created a grid of points on the virtual surface of the screen and sent a ray through each point. We embedded the ray-casting test into the game engine and replayed each demo file, thereby creating a record of objects visible to the player over the course of the game. Ray-casting, however, is a computationally costly technique. The more rays that are cast into a scene, the higher the cost. On the other hand, more rays lead to higher quality visual information. For this task a grid of 35 by 35 points proved to

Doors:
`through.*[door|one|that]`
`[door|one|that].*through`
Buttons
`[press|push|hit].*[button|one|that]`
`[button|one|that].*[press|push|hit]`

Figure 2: Regular expressions for instruction detection (*. matches any character sequence)

deliver reliable visibility results without being too computationally expensive.

Every 5 frames a "snapshot" of the visual field was taken that recorded which objects were visible, how many rays each object was hit by, and the horizontal angular deviation between a vector projected from the eye-point of the player into the center of the field of view, and a vector towards the center of mass of the object. The number of rays an object was hit by gives an indication of how much space in the field of view was occupied by this object.

Detecting the instructions

After recording the visibility information, we used regular expressions to identify instructions that were issued by the DG to get the DF to perform certain actions. We focused on instructions that requested two kinds of actions:

Passing through doors: To traverse from one room to another, the player has to pass through doors. The DG typically gives instructions like "*go through that door*" for this kind of action.

Activating buttons: To open or close cabinets the player has to activate buttons. The DG typically gives instructions like "*hit that button*" or "*push the button*" for this action.

We were interested in these two types of instructions because they make reference to objects in the world and, because the actions they request can be detected in the replays of the game, which opens up further research directions. We chose the regular expressions to use to find instructions so that they would detect explicit references to doors and buttons as well as indirect references such as expressions involving one anaphora. The expressions we used are shown in Figure 2. These patterns were chosen on the basis of an analysis of the first three dialogues in the corpus. We chose as the set of instructions all expressions that matched the pattern and contained not more than seven words. If the matched expression is smaller than seven words, the words following the expression up to a total length of seven words are added. Thus we set up a window of seven words around instructions to capture possible modifications around the detected instruction. Experience showed that seven words was a reasonable size, capturing most actual instructions without producing many false positives.

For each instruction we checked if the original annotations from the corpus contained a reference annotation and associated the annotated referent with the instruction. This formed the basis for the evaluation of the data and left us with a total of 318 instructions. These instructions together with the annotated referents formed the gold standard for the evaluation. Table 1 shows two sample instructions ,their annotated gold standard referents and the corresponding object visibility data.

Resolution of Referring Expressions

To evaluate the interpretation of referring expressions we had to bring together the data from the extracted corpus instructions and our object visibility data. We did this by synchronizing the visual record and instructions by their time stamp. For an instruction our algorithm performs the following steps in order to resolve the instruction's referent:

1. Extract which type of object (door or button) is referred to in the instruction by matching the instruction with the regular expressions detailed in Figure 2.
2. Collect all objects visible during the time covered by the instruction.
3. Filter out all objects of types incompatible to the instruction.
4. For each remaining object sum the number of ray hits for that object.
5. Rank the objects using a salience metric.
6. Return the object with the highest salience.

We developed four salience metrics of increasing sophistication by which the salience of objects can be ranked. These are as follows:

Baseline: As a baseline we took the stochastic probability that a randomly selecting process would pick out the correct referent from the set of visible objects and assigned this as the salience of each object.

Metric 1: This metric calculated the salience of each visible object by counting the total number of rays it was covered by. This metric works with the assumption that the object that is visually the largest is the intended referent. Figure 3 illustrates a screen that is overlayed with a grid of dots representing rays that are sent into the scene. In the scene two objects, **A** and **B**, are visible. Object **A** occupies a larger part of the screen than object **B** and receives more ray hits. Object **A** is therefore judged more salient by this metric.

Metric 2: The next metric weighted the number of ray hits for each object by the closeness of the objects to the assumed centre of attention. For this metric we assumed that the focus of attention would always be the centre of the screen. Figure 4 shows another grid of rays overlaying a screen. Here rays in the center receive a higher weighting, symbolized by the dots in the center being larger. Objects **A** and **B** occupy equally large areas of the screen. However, object **A** is in the center and so its hits are higher weighted than object **B**'s. Object **A** is therefore judged more salient.

Metric 3: This metric again weighted the number of rays hitting an object based on centrality but this time a mobile center of attention was used. If the 7 word window around the instruction contained the word "left", the

Instructions			Visibility Information
Start	Text	Annotated referent	Visible objects
00:03:34.214	*"through the door xxx and keep keep-"*	D8	{⟨D8, 4668,0.7⟩},{⟨D4, 1262,18.2⟩}
00:04:11.251	*"through the door uh that's closer xxx"*	D3	{⟨D3,1900,14.4⟩},{⟨D1,1461,-17.8⟩}

Table 1: Instructions and visibility information for instructions. Visible objects are represented as a triple of the name of the object, the ray count, and the angle of the object

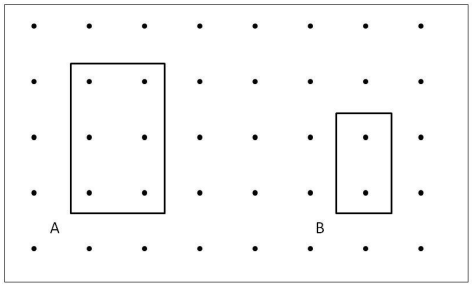

Figure 3: Salience based on the number of ray hits.

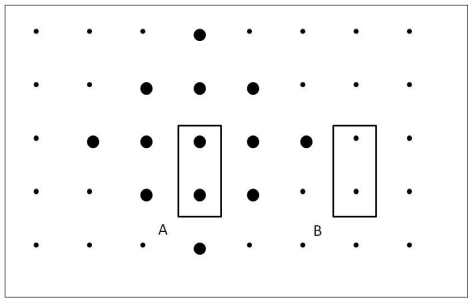

Figure 4: Salience based on weighted number of ray hits with central focus of attention.

center of attention would be shifted towards the left side of the screen. Conversely for the word "right". If neither "left" nor "right" was contained, this metric performs identically to Metric 2. Figure 5 illustrates this metric for an instruction containing the word "left". Objects **A** and **B** occupy equally large areas of the screen. However, object **A** is on the left of the screen where the center of attention is shifted and so its ray hits are weighted more highly than those hitting object **B**. Object **A** is therefore judged more salient.

The salience weighting was based on a linear drop off model as presented in (Kelleher and van Genabith 2004). The equation is presented in Equation 1. The weighting takes only horizontal deviation into account. The reason for this is that in the experimental setup all objects of a class occur at the same level, e.g. all buttons were at the same height and of the same size.

$$weight = 1 - \frac{|\gamma - \alpha|}{\alpha_m} \quad (1)$$

Angle α denotes the angle of the deviation between the central line of view of the player and the vector towards the

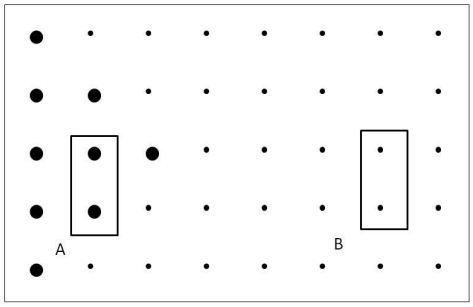

Figure 5: Salience based on weighted number of ray hits with left focussed center of attention.

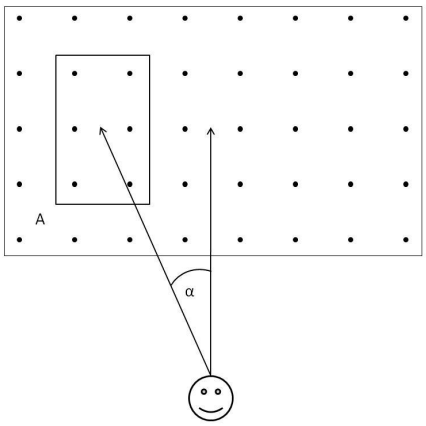

Figure 6: Angle between central line of view and vector towards an object.

object (this angle is illustrated in Figure 6). It can take values between -60 degrees and +60 degrees (for objects on the left and the right side of the field of view). α_m denotes the absolute maximum value α can take. γ denotes the angle of the central line of view of the player. It is set to 0 by default and set to -40 or +40 if the center of attention is shifted to the left or right respectively.

In the next section we describe how we evaluated our approach to resolving referring expressions and the various salience metrics.

Evaluation

To evaluate our approach to reference resolution we processed every instruction in the SCARE corpus using the algorithm described previously and the four salience metrics presented. The object deemed by our algorithm to be re-

	Total	Referent visible	Proportion of visible referents
Doors	195	166	85.1%
Buttons	123	104	84.6%
Both	318	270	83.6%

Table 2: Distribution of the detected instructions.

	Baseline	Metric 1	Metric 2	Metric 3
Doors	57.4%	55.0%	84.7%	85.6%
Buttons	52.0%	53.4%	79.4%	83.9%

Table 3: Proportion of correct predictions for each metric.

ferred to in each instruction was compared against the annotated gold standard referents to determine the accuracy of our approach. We did not consider instructions where the annotated referent was not visible. This was the case when the instruction contained a movement instruction that was to be executed before the referent enters the field of vision. Such an instruction may be "There should be a door behind you, go through that". Table 2 shows a breakdown of how many instructions were detected in total, for how many the annotated referent was visible during the instruction, and the proportion of instructions with visible referents.

The performance of our reference resolution approach is presented in Table 3, and in Figures 7 and 8. The first point to note from these results is that Metric 1 (based on the pure hit count) is relatively close to randomly picking out an object, i.e. the baseline metric. Metric 2 (based on a fixed center of attention), however, shows a clear improvement over Metric 1 and the baseline. This suggests that the center of the field of view is a workable approximation for the center of attention.

Finally, Metric 3 (based on a movable centre of attention) shows a slight increase over the Metric 2. Overall, the increase is not statistically significant with a two-tailed t-test giving a p-value of 0.84 for the doors domain, and 0.64 for the buttons domain. It should be noted though that the

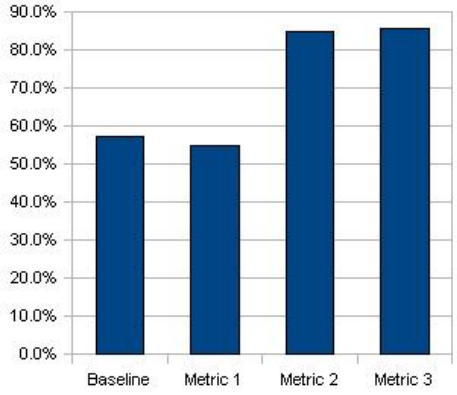

Figure 7: Proportion of correct predictions for the doors domain.

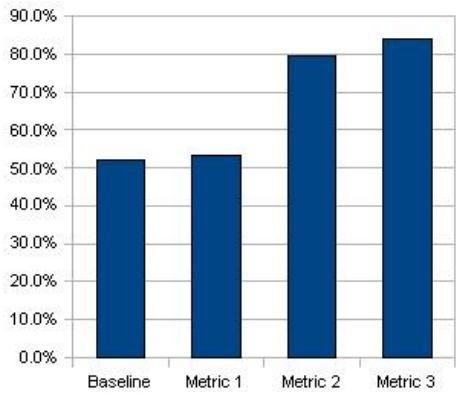

Figure 8: Proportion of correct predictions for the buttons domain.

value for the buttons domain is lower, possibly indicating a stronger effect of Metric 3 for the buttons domain.

We had expected a larger increase than that observed. The lack of this increase may be explained by an overall relatively small number of referring expressions using the chosen keywords (24 for the doors domain (13.0%) and 24 for the buttons domain(19.5%)). Also, directional keywords could appear within the instruction window without the intention to cause the assumed effect (e.g. "right there" in "Go through the door right there"). It is likely that this had a polluting effect.

Comparing the results for doors and buttons, it appears that the buttons domain shows slightly worse figures than the doors domain. This may be explained by the fact that buttons tend to be more closely grouped together than doors, thereby creating a higher possibility of confusion (this explanation is supported by the fact that the random approach worked slightly better for the doors domain and the higher proportion of expressions containing keywords). Figure 9 gives an overview of the perplexity of both domains. The horizontal axis gives different values for the size of the set of objects that were visible during the production of the instructions. The horizontal axis show how often each size of set occurred. The distribution for doors is skewed to the left, while the distribution for buttons is skewed to the right. The supports the impression that for buttons there is a higher possibility of confusion.

This may be related to the observation that the increase between metric 2 and 3 is slightly more noticeable for the buttons domain than for the door domain. This might indicate that users preferred to use spatial expressions to refer to buttons, which would be consistent with buttons being arranged in more complex configurations.

To more accurately determine the difference between metric 2 and 3, we selected only those cases where Metric 2 and 3 selected different referents and compared the results in isolation. For the doors domain, this gave us 9 cases and 6 for the buttons domain. The results for this analysis are presented in Table 4.

It is noticeable that for this set of cases, Metric 2 deliv-

	Baseline	Metric 1	Metric 2	Metric 3
Doors	40.3%	44.4%	33.3%	55.6%
Buttons	46.7%	50.0%	0 %	83.3%

Table 4: Proportion of correct predictions for each metric for cases where Metric 2 and 3 performed differently.

ers worse results than Metric 1 and the Baseline. Metric 3 however shows a clear improvement compared to the other metrics. This fits in well with our expectations: If the intended object is not in the center of the field of view (where Metric 2 focuses the attention), a decrease in performance is to be expected for Metric 2. At the same time, an increase for Metric 3 is to be expected. This indicates that shifting the center of attention has a positive effect, and that our simple approach for detecting when to shift is effective to a certain degree.

Conclusions & Future Work

We set out to evaluate if a simple, salience based mechanism can be used to resolve exophoric referring expressions in a situated dialogue. We evaluated this idea using the SCARE corpus augmented with object visibility information. The results show that at least for this type of dialogue this approach works reasonably well. The approach we developed is fairly accurate and computationally inexpensive. These properties make it particularly suitable for real-time human-robot interaction scenarios in which the human and the robot share a visual context, such as the tele-operated mobile robot scenario described in the introduction. We do not claim however, that this approach can be generalized to all types of dialogues.

For future work, we will explore the use of a more sophisticated method of determining instructions. If we can extract more information about spatial relations from the referring expressions, it may enable us to perform more accurate adjustment of the center of attention. The fact that a higher proportion of instructions containing our keywords in the buttons domain correlates with a higher improvement for Metric 3 against Metric 2 in comparison to the doors domain suggests that, while the simple use of keywords did have an effect, it probably was too blunt an approach. In a different direction, we extracted a set of abstract events such as the pushing of buttons from the available data. We will also correlate this data with the data about instructions and evaluate the level of ambiguity of referring expressions and reverse the approach taken in this work to attempt to learn rules for when to use instructions and what information to include in those instructions. Finally, we intend to implement our approach on an actual robot system such as that described previously.

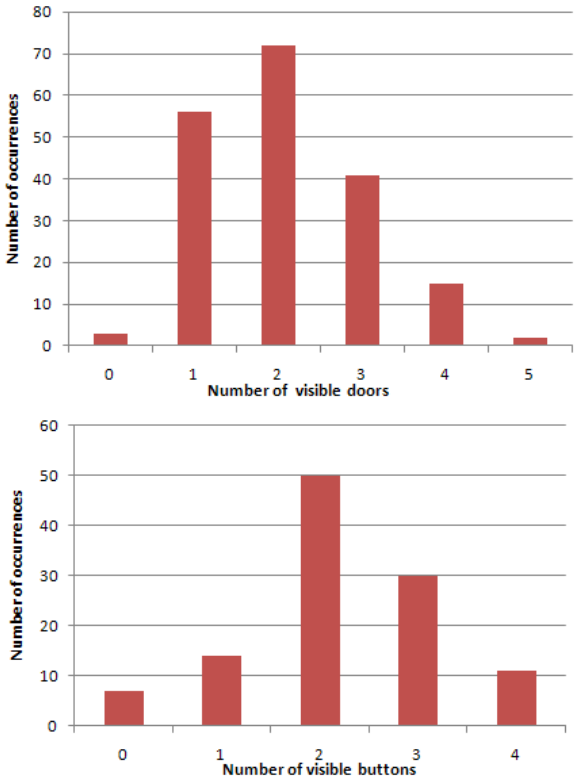

Figure 9: Distribution of number of objects visible during instructions.

References

Foley, J. D.; van Dam, A.; Feiner, S. K.; and Hughes, J. F. 1996. *Computer Graphics: Principles and Practice (2nd Edition)*. Addison Wesley.

Gorniak, P., and Roy, D. 2004. Grounded semantic composition for visual scenes. *Journal of Artificial Intelligence Research* 21:429–470.

Grosz, B. J.; Weinstein, S.; and Joshi, A. K. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics* 21:203–225.

Itti, L., and Koch, C. 2001. Computational modelling of visual attention. *Nature Reviews Neuroscience* 2(3):194–203.

Kelleher, J., and van Genabith, J. 2004. Visual salience and reference resolution in simulated 3-d environments. *Artificial Intelligence Review* 21(3).

Kelleher, J. D. 2006. Attention driven reference resolution in multimodal contexts. *Artif. Intell. Rev.* 25(1-2):21–35.

Kievit, L.; Piwek, P.; Beun, R.-J.; and Bunt, H. 2001. Multimodal cooperative resolution of referential expressions in the denk system.

Stoia, L.; Shockley, D. M.; Byron, D. K.; and Fosler-Lussier, E. 2008. Scare: A situated corpus with annotated referring expressions. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*.