Exploring the Behavior of Classic REG Algorithms in the Description of Characters in 3D Images *

Gonzalo Méndez and **Raquel Hervás** and **Susana Bautista** and **Adrián Rabadán** and **Teresa Rodríguez**
Facultad de Informática - Instituto de Tecnología del Conocimiento
Universidad Complutense de Madrid (Spain)
{gmendez,raquelhb,subautis}@fdi.ucm.es, {arabadan,teresaro}@ucm.es

Abstract

Describing people and characters can be very useful in different contexts, such as computational narrative or image description for the visually impaired. However, a review of the existing literature shows that the automatic generation of people descriptions has not received much attention. Our work focuses on the description of people in snapshots from a 3D environment. First, we have conducted a survey to identify the way in which people describe other people under different conditions. We have used the information extracted from this survey to design several Referring Expression Generation algorithms which produce similar results. We have evaluated these algorithms with users in order to identify which ones generate the best description for specific characters in different situations. The evaluation has shown that, in order to generate good descriptions, a combination of different algorithms has to be used depending on the features and situation of the person to be described.

1 Introduction

In every conversation, human beings refer to people, objects, places and situations, and we need to be able to describe them accurately so that the hearer knows who or what we are referring to. In order to be able to automatically create descriptions that can be useful in real life situations – such as generating descriptions for the visually impaired – where the complexity of the information needed to generate them is noteworthy, we first need to tackle specific aspects of these problems that bring light to the more general problem we intend to solve.

In this work we focus on the description of people in snapshots from a 3D environment, considering that feature extraction can be perfectly performed. Whereas most approaches to image description work with real world images, we have opted for 3D images because they allow us to easily manipulate the entities and their features in order to test different hypothesis, and we can create more complex situations to test our algorithms. In addition, we only focus on the description of people. Except for the TUNA corpus (Gatt et al., 2007; Deemter et al., 2012), which contained a set of close-up photographs of people, and the algorithms that used it in the TUNA Challenges (Gatt and Belz, 2010), to the best of our knowledge there are no works only focusing on people when describing visual images taken in real environments. The insights obtained in our work can improve the generation of descriptions for images where people are detected, as the examples presented by other studies show how references to people do not work in the same way as references to other entities do.

As a first step towards the implementation of a REG algorithm for describing people in 3D environments, we have explored the performance of classic REG algorithms for the task. We chose two well-known algorithms that can be easily configured depending on the the type of entities to be described: the Greedy (Dale, 1992) and the Incremental Algorithms (Dale and Reiter, 1995).

* This research is supported by the IDiLyCo project (TIN2015-66655-R) funded by the Spanish Ministry of Economy, Industry and Competitiveness.

Proceedings of The 10th International Natural Language Generation conference, pages 61–69,
Santiago de Compostela, Spain, September 4-7 2017.

As these algorithms require a predefined list of attributes that define the referent's appearance, we carried out a small study in order to determine the attributes that people include when describing people in real-life images (section 3). Then, we implemented the algorithms (and some variations) taking into account the obtained results (section 4), and asked people to judge the quality of their output when generating descriptions of characters in a 3D environment (section 5).

For both evaluations, we have taken an approach similar to the one in (Koller et al., 2010), which consists of an internet-based evaluation that allows for lower costs (it becomes unnecessary to summon a group of subjects to try out the system in a specific place). Users could easily access each survey using a link we provided, and they could do this at any time and from any place.

2 Related Work

A referring expression is a description created with the intention of distinguishing a certain object or person (*referent*) from a number of other objects or people (*distractors*). It must identify the referent unambiguously, effectively ruling out all the distractors. Therefore, any sentence that meets these criteria can be called a referring expression. However, not all of them can be considered equally good. It is usually considered that an effective referring expression should only contain information that the user knows or can easily perceive, and preferably information that is perceptually salient. In addition, overspecification could be desirable when extra information can help the listener to find the target more easily (Paraboni and van Deemter, 2014).

2.1 Classical Referring Expression Generation Algorithms

The task of Referring Expression Generation (REG) has been explored for over forty years (Krahmer and van Deemter, 2012). Although there are many other approaches to solve this problem (graph-based algorithms, constraint-based algorithms or description logics), in this work we have chosen to focus on two classical algorithms and the incorporation of relations into them. These algorithms are appropriate for our work because they are oriented to general purpose settings and therefore are easily configurable for different domains and situations.

The Greedy Algorithm (Dale, 1989; Dale, 1992) creates a reference by iteratively selecting the attribute with the highest discriminatory power which rules out most of the distractors. The algorithm continues working until there are no distractors left, or there are no attributes left (in which case the referring expression cannot successfully identify the referent). Since there is no backtracking, sometimes one of the attributes that has been included may become redundant as a result of the combination of other attributes used afterwards. For this reason the algorithm does not truly offer minimal referring expressions, but it does focus on the most salient properties of the referent.

The Incremental Algorithm (Reiter and Dale, 1992; Dale and Reiter, 1995) has been one of the most influential REG algorithms so far. It builds referring expressions incrementally, similarly to the Greedy Algorithm. The difference between the two is that the Incremental Algorithm has a list of attributes in a pre-established order, and in each iteration it picks the first one from the list that rules out at least one distractor. This method is more likely to lead to overspecification of the referring expression, since the algorithm does not allow backtracking. The order of the attributes is crucial, in this case the algorithm cannot select salient properties by itself, so this list should be chosen with care depending on the context or scene.

In addition to merely mentioning the properties of the referent, several algorithms have incorporated relations to other objects or people into their referring expressions, the first of which was the Relational Algorithm (Dale and Haddock, 1991). Since then, relations have been incorporated into other algorithms, but they are very often considered inferior to properties belonging to the referent itself, and are used only as a last resort when its attributes are not enough to distinguish it (Krahmer and Theune, 2002). However, there is also research that proves that people tend to use relations in their descriptions even when they are not necessary (Viethen and Dale, 2008). Works like the ones by Kelleher and Kruijff (2005) deal with the determination of the best landmarks to use in a referring expression depending on context.

2.2 Automatic Description of Visual Information

The automatic generation of image descriptions is a problem that has received a large amount of interest in recent years from both computer vision and natural language generation communities.

An extensive survey on this topic can be found in (Bernardi et al., 2016). The authors divide the existing approaches into two main groups based on the models used. *Direct generation models* follow a classical pipeline: they first extract image information in terms of entities, relations between them, etc., and then this information is used by a natural language generation algorithm to generate the final image description. *Retrieval models* attack the problem by searching for images that are similar to the one to be described and then building the final description based on the descriptions of the retrieved images. Because our work consists in the description of characters in an interactive setting, we are more interested in direct generation models where a previously available database of image and descriptions is not required.

Although direct generation models have the advantage of being able to produce novel descriptions without relying on a previously existing corpus of descriptions, they rely heavily on the quality of the conceptual information extracted from the original image. In order to tackle this issue, some authors have started to separate both problems and study the generation of image descriptions assuming that visual image recognizers have already achieved close to perfection identification of information in images (Elliott and Keller, 2013; Yatskar et al., 2014; Wang and Gaizauskas, 2015).

3 Identification of Features Used in Descriptions

We conducted a survey in order to identify what features are relevant for individuals when they have to describe other people. A total of 71 evaluators took part in this survey. They were presented with photographs taken in our university canteen which contained a high number of people (an example can be seen in Figure 1) and they had to complete two sets of tasks.

3.1 Part 1: Identifying People

In this part of the survey, the participants were provided with four pictures of the canteen, each of them accompanied by a description, and they were asked to *"Find the person described at the top of the screen"*.

In the first scene, the participants were asked to identify a boy with a black t-shirt. In this picture, four boys were dressed in black, but two of them were wearing coats instead of t-shirts. Any of the other two were considered as a correct answer. 32% of the people chose a boy wearing a black coat, who was the most visible person in the scene and the closest one to the observer. 49% chose either of the two boys wearing a black t-shirt (28% and 21%, respectively) and 7% did not know the answer. From these answers we concluded that people are more likely to notice someone who is closer to them, and that the color of a person's clothes is more important that the type of the clothes.

In the second scene, the participants were asked to identify a boy leaning against a wall. We intended to find out if it would be easier for the participants to identify a person when they are very close to an important area in the room. 94% chose the right individual. He is at the edge of the photo and he is not very visible, but he is the only one leaning on the wall. The conclusion in this scene is that, since the wall is an important part of the room, people's eyes are drawn to it quickly, making it easy for them to find the person they are looking for.

In the third scene, the participants were asked to identify a person sitting next to a window. This time, as well as choosing a person that is next to an important area of the room, we picked someone who was further away from the user, to see if this had any effect on the participants reactions. 96% of the participants chose the right boy. By mentioning a relevant element such as the window, people's attention seem to automatically go towards that area and ignore the rest of the picture, so it is easier for them to find the person who fits the description.

In the fourth scene, the participants were asked to identify a girl with black hair. We chose a person furthest away from the viewer, and we decided to pick one of the only two girls with dark hair in the whole photograph. 69% of the people chose the

Figure 1: Sample scene used in the first study

correct girl, even though, out of all the girls, she was the one that was the furthest away from the observer. 23% chose a girl with dark (not black) hair, closer to the observer than the right girl. Two people chose a blond girl at the front of the photo, and two more did not know the answer. From these answers we can see that people tend to focus on what they see first. For this reason, it may be a good idea to provide more details than necessary when describing a person that is further away.

3.2 Part 2: Describing People

In this part of the survey, participants were provided with several pictures and were asked to *"Describe the person number N"* (see Figure 1).

In the first scene, the participants had to describe a boy working with his laptop. 66% of the participants mentioned his posture in some way (e.g. leaning on the table, working with his laptop), and 36% mentioned his clothes. We can conclude that, in this case, since the referent was in a very particular posture (hands crossed beneath his chin and looking at his laptop), the users have a tendency to include this as the main part of their description. There is only one other person in the photograph with a laptop, and nobody else visible with their hands under their chin. For this reason his posture stands out as a very descriptive feature.

In the second scene, they had to describe a waitress of the canteen. Overall, 59% of the participants mentioned her clothes, and 41% mentioned her profession. We can infer that, when someone is recognizable by their type, this can be descriptive enough and we may not need to mention anything else.

In the third scene, the target boy was barely visible. 8% of the participants gave an exhaustive description of everything they could see, but a lot of people described him by his clothes (53%) even though there are other boys close in the picture who are wearing clothes of a similar description (white t-shirt with dark details). Even when there are several people in a scene wearing similar clothes, people of-

ten tend to include information about those clothes in their description.

In the fourth scene, 21% of the participants described the target person as the boy with the red shirt, and did not mention anything else, even though there is another boy that could also fit in that description. A few people also noticed his posture (24%) and the fact that he is within a group of people. This reinforces what we concluded in the first part of the survey: when people see someone who fits a description, they do not look any further to check if that description may apply to someone else.

In the fifth scene, the target boy is sitting with a group of friends and is wearing a red shirt, so his description might be very similar to one of his friends. This time, 31% of the people described his posture as well as his clothes, and said that he is talking to the boy next to him. 14% of the participants described only his clothes, but they mentioned that his top has long sleeves, in contrast to his friends t-shirt. Even when the color of their clothes alone is not enough to distinguish a person, if it stands out enough, users tend to mention only that.

In the sixth scene, the target boy's face is not visible, the color of his clothes does not stand out, and there seems to be nothing particularly eye-catching about him. In this case, 73% of the people described his posture (he is sitting facing away from the observer), and most mentioned that he is sitting next to a girl. Some even described the girl's clothes, because they stand out more than his. Here we can see that when a person does not stand out very much, people tend to notice something nearby that stands out more (in this case the girl he is sitting with, but it could also be a window, a door or an object like a laptop, as seen in previous scenes).

3.3 Results

This study provided us with two important insights. The first one was that distance (from the viewer and to landmarks) influences the identification of referents. We could observe in the survey that the test subjects sometimes focused on the people who were nearer to them in the scene, and if a distractor looked similar to the referent, even if not all the attributes in the description matched, they would settle for this distractor. It also seemed that referring expressions that include information about nearby objects or people were easier to understand.

The second insight obtained from the study was a list of preferred attributes when describing people in crowded environments. The type of the person (e.g. boy, girl, waitress) was mentioned very often, in an average of 73.11% of the description. The next most used attribute was the colour of the top garment, and the last attribute which stood out was posture (used on average in 57.31% of the descriptions). Interestingly, the test subjects only mentioned important areas of the room 13.91% of the time, and described nearby people a little more often, 17.21% of the time.

Finally, based on the results we have obtained, we have seen that, rather than giving the shortest and most efficient description possible, people often give more information than is needed. This makes it easier for us to find the right person quickly.

4 Implementation of Classic Algorithms

With the results obtained from the previous study we could implement algorithms that do not choose the included attributes arbitrarily, but based on the opinions of real test subjects. Some of the attributes mentioned by the evaluators were not used because they are either too subjective (attitude, personality, age, height, weight) or cannot be appreciated in an image (shoe type and colour). The resulting prioritized list of attributes for the Greedy and Incremental Algorithms is therefore the following (from most to least priority): (1) type; (2) top colour; (3) posture; (4) beard; (5) hair colour; (6) top type; (7) hair type/length; (8) bottom colour; (9) bottom type.

In light of the results of the previous survey, we decided to include information about whether the referent is close to or far away from the observer in order to distinguish the referent faster. Considering the size of the room used in the scenes, we divided the space into two halves. The distance in the environment was measured from the observer to each character, so the character who was furthest away dictated the maximum distance that would be considered, and this would be divided by two to create a halfway division. Every character who was between the observer and the division would be considered near, and the rest would be considered far. Since this is not strictly a physical attribute of the referent

it was not included in the Greedy Algorithm. For example, distance could potentially discard a large amount of people in the Greedy Algorithm while not being clearly visible to the observer if they are all in a group but some of them are standing further back. Distance was mentioned only in the Incremental Algorithm and it was added at the end of the description.

In addition to the Greedy and Incremental Algorithms, we also included the Exhaustive Algorithm as a baseline, which offers a full description of the referent including all its features. This last type of referring expression can be overspecified or non-distinguishing, so it is not ideal for describing the referent. The sentence structure in the Exhaustive Algorithm, taking into account the previously prioritized list of attributes, is:

> *The Type with HairType HairColour hair [and a Beard], with the TopColour TopType and BottomColour BottomType.*

Finally, in order to take into account objects or people near the referent, we implemented two relational algorithms to be used in combination with the three previously mentioned. Therefore, the referent was described using one of the three previous algorithms and additional information about relevant people or objects was included in case there was any.

The Nearby Objects Algorithm checks if there are any significant areas or objects near the referent and mentions the closest one. The referent can be described using either of the three basic algorithms, which leaves us with three different versions of the Nearby Objects Algorithm.

The Nearby People Algorithm works in a similar way, but the distance required to consider a person next to the referent is a little longer than in the previous case, since people tend to keep slightly further away from other people than from objects (although this distance is known to be culture dependent).

Therefore, the final algorithms were the Exhaustive Algorithm (EA), the Incremental Algorithm (IA), the Greedy Algorithm (GA), Nearby Objects with Exhaustive Algorithm (NOEA), Nearby Objects with Incremental Algorithm (NOIA), Nearby Objects with Greedy Algorithm (NOGA), Nearby People with Exhaustive Algorithm (NPEA), Nearby People with Incremental Algorithm (NPIA) and Nearby People with Greedy Algorithm (NPGA). Out of these nine algorithms, NOEA and NPEA have been excluded from the evaluation, since the EA algorithm was included only as a baseline and some preliminary tests pointed out that the descriptions provided by NOEA and NPEA algorithms did not improve the ones provided by the other algorithms. On the contrary, overspecification decreased the quality of these descriptions.

5 Evaluation of Classic Algorithms

After implementing all the algorithms, we tested them in order to find out if there was one that worked better than the rest in all situations or which one worked best depending on the situation.

In this survey, we showed the participants snapshots of the university canteen taken in a 3D virtual environment built using the Unity 3D engine. The characters' clothes and postures were modified to imitate the ones in the pictures (see Figures 1 and 2). The use of a 3D environment aids in the personalization of the scene, facilitating experiments in which any number of people and objects can be represented. This way we were also able to appreciate the differences between the descriptions given for a photograph of a real scene, and a scene developed in a 3D virtual environment.

A total of fifty-two participants completed this survey: 54% were women and 46% were men; most of them (67%) were between eighteen and thirty years old, 17% were between thirty and forty years old, 4% were under eighteen, and 12% were over forty.

The structure of the survey and the order of the questions were carefully planned so they did not influence the users' opinions. We wanted them to offer their own descriptions first, before reading and judging the descriptions generated by the algorithms. We have also considered the effort and amount of time that they will have to spend on the survey, so they will not be tempted to leave it unfinished and we can get as many answers as possible.

Because both the way in which the test subjects describe someone in the 3D scene and their ability to recognize the target character given a referring expression were intended to be analyzed, this survey was divided in two parts.

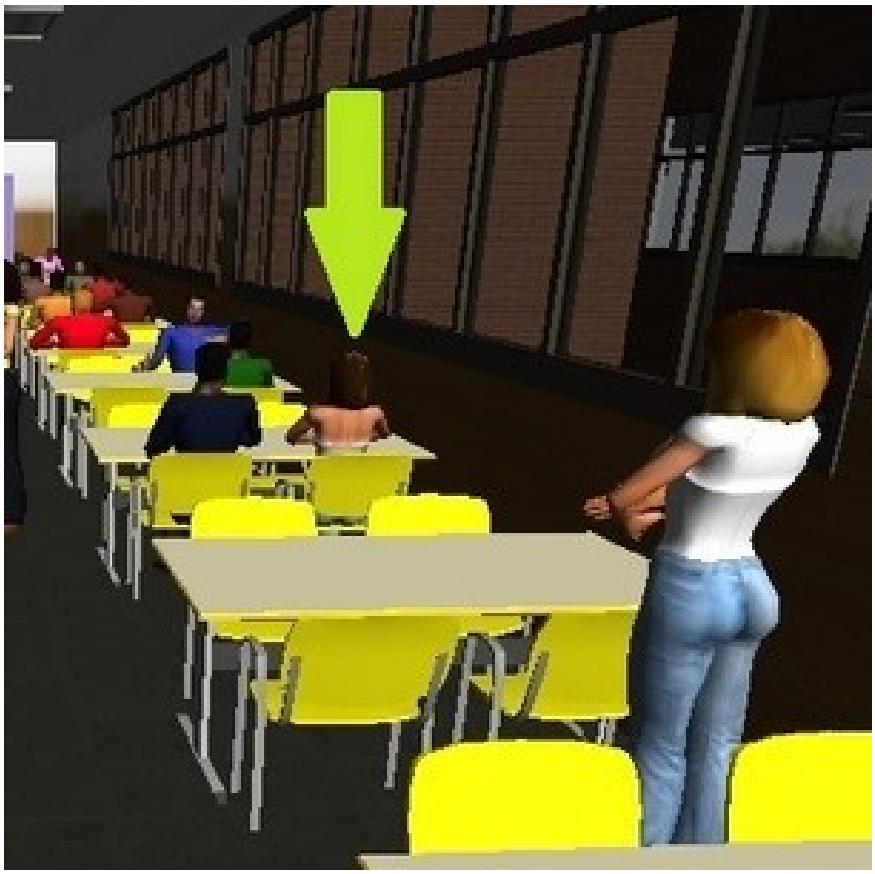

Figure 2: Sample scene used in the second study

In the first part, each test subject was asked to describe a certain person from three different scenes. The goal was to examine whether their answers would be very different when faced with a 3D environment as opposed to photographs.

The results show that type, top color, top type and posture still were the most used attributes, and they were mentioned even more often than during the first survey. The difference between inclusion of the color of the top garment and its type increased slightly, confirming that the color is a more salient attribute. The inclusion of nearby people and nearby objects approximately doubled in both cases, possibly due to the simplified representation of the room and the characters. In the case of the nearby people, we could see that it is not always the closest person that gets mentioned, but the person that stands out the most among the closest ones.

The use of hair color decreased slightly, and hair type/length was rarely used, possibly because there were not many variations of hairstyles in the scene. Even though two of the referents had a beard, the test subjects only mentioned it in 5.77% of the descriptions, much less than in the first survey and contrary to our hypothesis. This may be either due to the quality of the characters used or simply because the beard is not a very salient attribute.

Overall, the results showed a similar order in the preferred attributes, with relations to large areas and other people gaining more importance, and small details being used less.

For the second part of the survey, the test subjects were shown four scenes (see Figure 2), each of them with several referring expressions for one referent, created and linguistically realized by our algorithms. Then, they had to rate each of the descriptions on a five point Likert scale, the lowest value being “very bad” and the highest “very good”. Not all the algorithms were rated in all the scenes, either because they did not provide any useful information (e.g. there were no nearby objects in scene 1, so NOGA and NOIA were discarded), or because they gener-

	EA	GA	IA	NOGA	NOIA	NPGA	NPIA
Scene 1	2.647	2.019	3.372	-	-	-	4.673
Scene 2	1.901	1.843	2.176	2.941	2.960	2.285	3.115
Scene 3	2.940	3.784	-	4.140	3.882	-	-
Scene 4	-	2.411	2.500	4.215	-	3.200	-

Table 1: Average scores obtained by the algorithms

ated results equivalent to those of other algorithms (e.g. NOIA and NOGA in scene 4). The following example shows the descriptions generated for Figure 2:

- Greedy (GA): *"The girl sitting down"*
- Incremental (IA): *"The girl in the white tank top who is sitting down. She is near."*
- Exhaustive (EA): *"The girl with medium length brown hair, with the white tank top and blue trousers."*
- Nearby Objects with Greedy (NOGA): *"The girl sitting down near the window."*
- Nearby Objects with Incremental (NOIA): *"The girl in the white tank top who is sitting down. She is near. She is near the window."*
- Nearby People with Greedy (NPGA): *"The girl sitting down next to the boy in the dark blue sweater."*
- Nearby People with Incremental (NPIA): *"The girl in the white tank top who is sitting down. She is near. She is next to the boy in the dark blue sweater."*

The obtained results are shown in Table 1, where the average score for the descriptions generated by the algorithms in each of the four scenes are shown. Relational algorithms have proved to have very high ratings. This suggests that, at least for the particular scenes and situations shown to the participants, relational algorithms which include nearby people or objects can be very useful if there are distractors or objects that stand out and can be related to the intended referent.

The results from the second survey also showed that users do not benefit from the inclusion of the beard or information about the referent's bottom garment or shoes, so we eliminated these from the attributes list. In scenarios in which people wear very unusual clothing this may not be a correct decision, but since we are working with characters with casual attire, the bottom half of their clothes are not different enough from each other to stand out. Additionally, many characters are sitting down or are partially covered and some parts of their clothes are often not visible to the observer.

6 Conclusions and Future Work

In the present work, we have described a user driven approach to automatically generate character descriptions in 3D environments. We have conducted two different surveys that have allowed us to identify, on the one hand, what attributes are more relevant for people when they describe another person, and on the other hand, what kind of description they understand better depending on the specific features and situation of the target subject of the description.

In our aim to build an algorithm that describes people in different, static, situations, the next step we must take is to design an strategy that, for a given scene, identifies the relevant features of the subject to be described and selects the most appropriate algorithm, among the studied ones, to generate a suitable description of this person.

In the long run, we intend to generate descriptions in closer to real life situations, where both the observer and the elements of the scene, either objects or people, can move and change, so that these changes have to be taken into account in order to modify the contents of the description in real time.

References

Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. 2016. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *Journal of Artificial Intelligence Research*, 55(1):409–442, January.

Robert Dale and Nicholas Haddock. 1991. Generating referring expressions involving relations. In *Proceedings of the 5th Conference of the European Chapter of the Association for Computational Linguistics*, pages 161–166, Germany.

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

Robert Dale. 1989. Cooking up referring expressions. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, pages 68–75, University of British Columbia, Vancouver, BC, Canada.

Robert Dale. 1992. *Generating Referring Expressions: Constructing Descriptions in a Domain of Objects and Processes*. MIT Press, Cambridge, Cambridge, MA, USA.

Kees van Deemter, Albert Gatt, Ielka van der Sluis, and Richard Power. 2012. Generation of referring expressions: Assessing the incremental algorithm. *Cognitive Science*, 36(5):799–836.

Desmond Elliott and Frank Keller. 2013. Image description using visual dependency representations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP-13)*, pages 1292–1302. Association for Computational Linguistics.

Albert Gatt and Anja Belz, 2010. *Introducing Shared Tasks to NLG: The TUNA Shared Task Evaluation Challenges*, pages 264–293. Springer Berlin Heidelberg, Berlin, Heidelberg.

Albert Gatt, Ielka van der Sluis, and Kees van Deemter. 2007. Evaluating algorithms for the generation of referring expressions using a balanced corpus. In *Proceedings of the Eleventh European Workshop on Natural Language Generation*, ENLG '07, pages 49–56, Stroudsburg, PA, USA. Association for Computational Linguistics.

John Kelleher and Geert-Jan Kruijff. 2005. A context-dependent algorithm for generating locative expressions in physically situated environments. In *Proceedings of the 10th European Workshop on Natural Language Generation*, pages 68–74, Aberdeen, UK.

Alexander Koller, Kristina Striegnitz, Donna Byron, Justine Cassell, Robert Dale, Johanna Moore, and Jon Oberlander. 2010. The first challenge on generating instructions in virtual environments. In E. Krahmer and M. Theune, editors, *Empirical Methods in Natural Language Generation*, volume 5790 of *LNCS*, pages 337–361. Springer.

Emiel Krahmer and Mariet Theune. 2002. Efficient context-sensitive generation of referring expressions. In Kees van Deemter and Rodger Kibble, editors, *Information Sharing: Reference and Presupposition in Language Generation and Interpretation*, pages 223–264. CSLI Publications, Stanford CA, USA.

Emiel Krahmer and Kees van Deemter. 2012. Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218.

Ivandr Paraboni and Kees van Deemter. 2014. Reference and the facilitation of search in spatial domains. *Language, Cognition and Neuroscience*, 29(8):1002–1017.

Ehud Reiter and Robert Dale. 1992. A fast algorithm for the generation of referring expressions. In *Proceedings of the 14th International Conference on Computational Linguistics*, pages 232–238, Nantes, France.

Jette Viethen and Robert Dale. 2008. The use of spatial relations in referring expression generation. In *Proceedings of the 5th International Conference on Natural Language Generation*, pages 59–67, Salt Fork, OH, USA.

J.K. Wang and R. Gaizauskas. 2015. Generating image descriptions with gold standard visual inputs: Motivation, evaluation and baselines. In *15th European Workshop on Natural Language Generation (ENLG)*, pages 117–126. Association for Computational Linguistics.

Mark Yatskar, Michel Galley, Lucy Vanderwende, and Luke Zettlemoyer. 2014. See no evil, say no evil: Description generation from densely labeled images. In *Proceedings of the Third Joint Conference on Lexical and Computational Semantics (*SEM 2014)*, pages 110–120. Association for Computational Linguistics.