# GRE3D7: A Corpus of Distinguishing Descriptions for Objects in Visual Scenes

**Jette Viethen**[1,2]
jette.viethen@mq.edu.au

[1]TiCC
University of Tilburg
Tilburg, The Netherlands

**Robert Dale**[2]
robert.dale@mq.edu.au

[2]Centre for Language Technology
Macquarie University
Sydney, Australia

## Abstract

Recent years have seen a trend towards empirically motivated and more data-driven approaches in the field of referring expression generation (REG). Much of this work has focussed on initial reference to objects in visual scenes. While this scenario of use is one of the strongest contenders for real-world applications of referring expression generation, existing data sets still only embody very simple stimulus scenes. To move this research forward, we require data sets built around increasingly complex scenes, and we need much larger data sets to accommodate their higher dimensionality. To control the complexity, we also need to adopt a hypothesis-driven approach to scene design. In this paper, we describe GRE3D7, the largest corpus of human-produced distinguishing descriptions available to date, discuss the hypotheses that underlie its design, and offer a number of analyses of the 4480 descriptions it contains.

## 1 Introduction

Whenever we engage in any form of discourse we need to find a way to describe to our readers or listeners the entities that we are talking or writing about. This act of referring to real-world entities is one of the central tasks in human language production. Of course, it is also central when a machine is charged with the task of generating natural language, which makes referring expression generation (REG) an important subtask in any natural language generation (NLG) system.

It is therefore not surprising that REG has attracted a great deal of attention from the NLG community over the past three decades. A key factor that has led to the popularity of REG is the widespread agreement that the central task involved is *content selection*: choosing those attributes of a target referent that best distinguish it from other distractor entities around it (Dale and Reiter, 1995; van Deemter, 2000; Gardent, 2002; Krahmer et al., 2003; Horacek, 2003; van der Sluis, 2005; Kelleher and Kruijff, 2006; Gatt, 2007; Viethen and Dale, 2008).

Recent work in particular has concentrated on the development of algorithms concerned with the generation of context-free identifying descriptions of objects, as emphasised by three shared-task evaluation competitions (STECs) targeting this particular problem (Belz and Gatt, 2007; Gatt et al., 2008; Gatt et al., 2009). Referring expressions of this kind are often referred to as *distinguishing descriptions*. We are still far from a full understanding of how such descriptions should best be generated. Much work remains to be done before many issues, such as, for example, the generation of relational descriptions and over-specified descriptions or the number of the surrounding objects to be taken into account in visual settings, can be considered resolved.

Although many authors have explicitly or implicitly acknowledged the importance of generating referring expressions that sound natural (Dale, 1989; Dale and Reiter, 1995; Gardent et al., 2004; Horacek, 2004; van der Sluis and Krahmer, 2004; Kelleher and Kruijff, 2006; Gatt, 2007; Gatt et al., 2007), much of the original work in REG was neither developed based on empirical evidence about
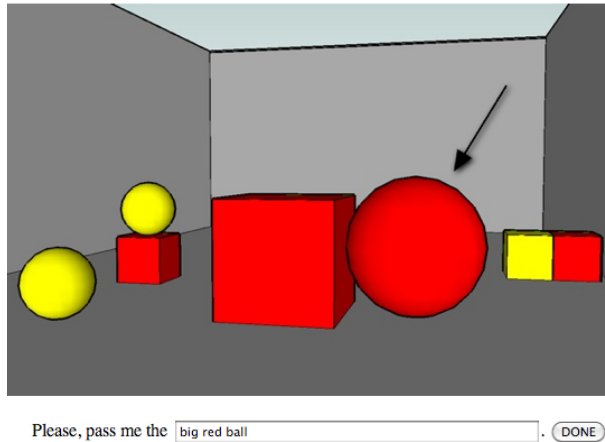
12

Please, pass me the [big red ball] . (DONE)

Figure 1: **The screen showing the first stimulus scene.**

how humans refer, nor evaluated against human-produced referring expressions. The REG STECs on the task of content determination form part of a recent trend towards more data-oriented development and evaluation of REG algorithms that responds directly to this concern (Gupta and Stent, 2005; Jordan and Walker, 2005; Gatt et al., 2007; Viethen et al., 2010; Belz and Gatt, 2007; Gatt et al., 2008; Gatt et al., 2009).

However, the existing data sets used in these experiments involve very simple and usually abstract visual displays of objects rather than coherent scenes. This is a reasonable starting point for bootstrapping research; but if we want to develop algorithms that can be used in real-world scenarios, we ultimately need to work with scenes which are much more realistic. At the same time, given the non-deterministic nature of choice in the production of natural language, corpora based on these scenes need to be very large, and should ideally contain referring expressions from as many different speakers as possible for each target referent in each referential scenario. The choice of stimuli and data collection procedure should provide a controlled environment that allows the isolation of a small number of factors influencing the choices that have to be made by the participants, in order to facilitate the replication of the same controlled environment for REG algorithms attempting the same reference task in an evaluation situation. The way forward, we believe, is to build a succession of corpora with incrementally more complex scenes.

In this paper, we describe the design of a data collection experiment for distinguishing descriptions and give an overview of the resulting corpus, which is, at 4480 instances, the largest corpus of distinguishing descriptions developed to date.[1] Consistent with the common focus on initial reference in visual scenes, we used visual stimuli containing a small number of simple objects (cubes and balls) in a 3D scene, similar to our much smaller GRE3D3 Corpus (Viethen and Dale, 2008), and elicited individual descriptions in the absence of a complicating preceding discourse. Additionally, we introduced factors that allow the study of the use of spatial relations in referring expressions by creating stimulus scenes that encourage the use of relations between objects, but do not require them. Most existing REG algorithms that can make use spatial relations between objects only do so if no distinguishing description can be found otherwise (Dale and Haddock, 1991; Gardent, 2002; Krahmer and Theune, 2002; van der Sluis and Krahmer, 2005; Kelleher and Kruijff, 2006), often based on the argument that mentioning two entities imposes a higher cognitive load than referring to only one entity. We are interested in investigating in how far this behaviour corresponds to the human use of spatial relations in distinguishing descriptions, as well as testing a number of concrete hypotheses about the factors that might lead people to use spatial relations.

## 2 Stimulus Design

The stimulus scenes used for the GRE3D7 corpus are three-dimensional scenes containing only simple geometric shapes, created in Google SketchUp. Each stimulus scene contains seven objects; these are grouped into three pairs of two and one single object. The target object is always part of one of the pairs and the second object of that pair is what we call the *landmark* object in these scenes. We attempted to place the target–landmark pair as close to the centre of the scene as possible to encourage the use of the target's direct object properties and its spatial relations to other objects, rather than its overall location in the scene, as in *in the left*. The other two object pairs were placed slightly further back to

---

[1]The corpus is available for download online at www.clt.mq.edu.au/research/projects/gre3d7.

the left and right of the target–landmark pair, and the single object was always placed in the far right or the far left of the scene. Objects were of one of two types (ball or cube) and otherwise distinguishable by their size and colour. Each object could be either large or small, and in each scene we used only two colours. Figure 1 shows a close-up of one of the scenes as presented to the subjects, and Figure 2 shows the complete set of stimulus scenes.

The design of the stimulus scenes was based on a number of hypotheses about the factors that might influence people's use of spatial relations to the landmark object. The two main hypotheses are concerned with the influence of the landmark object's size on its visual salience and the likelihood of the target–landmark relation being used in a referring expression:

**Hypothesis 1:** A large landmark is more salient than a small one because it occupies more of the visual space of a scene. Therefore, a large landmark is more likely to be mentioned in a referring expression via its spatial relation to the target referent than a small landmark.

**Hypothesis 2:** A landmark that shares its size with a number of other objects in the scene is less salient than one that is unique in size. Therefore, a landmark with unique size is more likely to be mentioned in a referring expression via its spatial relation to the target referent than a landmark with a common size.

Hypotheses 1 and 2 are concerned with the landmark's overall salience in the scene, or what is usually called *bottom-up* salience in the literature on visual attention (cf., Yantis, 1998). A second consideration that might influence the use of relations is the *top-down* salience of the target and landmark objects, as determined by the task the participants are performing. At the time when the landmark's visual salience is taken into account, the participants are focusing their attention on the target object. As the landmark is the closest object to the target, it is likely that the difference or similarity between these two objects plays a particularly important role in the decision whether to include the relation between them or not. Two conflicting hypotheses can be formulated here:

**Hypothesis 3:** The difference between the landmark and the target object impacts on the visual salience of the landmark because it impacts on the landmark's overall uniqueness in the scene. Therefore, a landmark that is visually different from the target is more likely to be included in a referring expression than one that looks similar to the target.

**Hypothesis 4:** The more similar the landmark and target objects are, the more they appear as one visual unit rather than two separate objects. If they are perceived and conceptualised as a visual unit, they are more likely to be mentioned together. Therefore, the more similar the landmark is to the target, the more likely it is to be included in a referring expression.

The fifth hypothesis that this experiment is designed to test concerns the preference that participants in psycholinguistic work have shown for vertical relations over horizontal ones (Lyons, 1977; Bryant et al., 1992; Gapp, 1995; Bryant et al., 2000; Landau, 2003; Arts, 2004; Tenbrink, 2004). To make sure that the landmark is never obscured by the target object, we use lateral relations rather than frontal ones in this experiment.

**Hypothesis 5:** A target placed on top of a landmark object is more likely to be described in terms of its spatial relation to the landmark than a target that is sitting directly adjacent to the left or right of the landmark.

We report the results of putting these five hypotheses to the test in Section 5.4. To be able to perform these tests systematically, the experiment was designed as a $2 \times 2 \times 2 \times 2 \times 2$ grid with the following five variables:

- LM_Size: the landmark is either large or small. [Large/Small]

- LM_Size_Rare: the size of the landmark is either a common size in the scene, or it is as rare as possible, and possibly unique. If it is common and the landmark is large, it shares its size with two of the objects; if it is small, with three. These numbers are not the same because in each scene in which the landmark

14

size was common, three objects were large and four small. In +LM_Size_Rare scenes that are also +TG_Size = LM_Size, the landmark shares size only with the target. Only if the scene is −TG_Size = LM_Size can the landmark's size be truly unique in the scene. [+/−]

- TG_Size = LM_Size: target and landmark are either the same size or different. [+/−]

- TG_Col = LM_Col: The target and the landmark are either of the same colour or different in colour. [+/−]

- Relation: The relation between the target and the landmark is either vertical (the target is on top of the landmark) or lateral, in which case the target is placed directly to the left or right of the landmark. [Vertical/Lateral]

This resulted in 32 experimental conditions. We created one stimulus scene for each of these conditions. We then split the stimuli into two trial sets along the factor TG_Size = LM_Size, so that this variable became a between-participant factor, while the other four are within-participant factors.

We followed a number of other criteria for the design of the stimulus scenes to ensure maximum experimental control over the factors influencing the content of the referring expressions provided by our participants:

**Target uniqueness:** The target was always distinguishable in terms of its inherent properties alone,[2] which means that the relation to the landmark or other external properties, such as the location in the scene, were never necessary to fully distinguish the target from all other objects in the scene.

**Landmark uniqueness:** As the target, the landmark was always distinguishable in terms of its inherent properties alone.

**Colour balance:** Each scene followed one of two colour schemes: either blue–green or red–yellow. The colour schemes were distributed in a balanced way across the five experimental variables, so that

half of the scenes in each condition were blue–green and the other half red–yellow. The colour scheme was not expected to have an influence on the content of the referring expressions people produced. In each scene, four objects were of one colour of the colour scheme for this scene and three had the other colour.

**Relation balance:** The relation between the target and the object was never unique. One of the two other object pairs in each scene was arranged in the same spatial relation as the target–landmark pair and the third pair had the other relation. However, the objects in the pair with the same relation were never of the same types as the target and landmark, so that a description containing the type of the target, a relation to the landmark and the type of the landmark was always fully distinguishing.

**Constant landmark and target types:** The landmark was always a cube, in order to avoid scenes where the target would have to be balanced on top of a ball, which might look unnatural. The target was always a ball to make sure that the similarity in type between these two objects was always constant.

**No obscured objects:** The objects were placed in the scenes in such a way that no object occluded any other. In particular, as mentioned above, there were no frontal relations within the object pairs, to avoid larger objects obscuring smaller ones completely or to a large degree.

Figure 2 shows the $2 \times 2 \times 2 \times 2 \times 2$ grid of the 32 stimuli scenes. Scenes 1–16, shown on a green background, constitute Trial Set 1, and Scenes 17–32, shown on a blue background, constitute Trial Set 2.

## 3 Procedure and Participants

The data gathering experiment was designed as a self-paced on-line language production study. Participants visited a website, where they first saw an introductory page with a set of simple instructions and a sample stimulus scene. Each participant was assigned one of the two trial sets containing 16 stimulus scenes each. After the instruction page, the scenes were presented consecutively in an order that was randomised for every participant. Below each scene, the participants had to complete the sentence

Figure 2: **The 32 stimulus scenes for GRE3D7**: The left half constitutes Trial Set 1 and the right half is Trial Set 2.

*Please pick up the . . .* in a text box before clicking a button labelled 'DONE' to move on to the next scene, as shown in Figure 1. The task was to describe the target referent in the scene (marked by a grey arrow) in a way that would enable a friend looking at the same scene to pick it out from the other objects. To encourage the use of fully distinguishing descriptions, participants were told that they had only one chance at describing the object.

Before each of the 16 stimulus scenes, the participants were shown a filler scene, which means each participant had to describe 32 scenes in total. The main motivation for using filler scenes was to minimise the decline in relation use over time, which might otherwise happen if participants realised that relations were never necessary.

The filler scenes were also designed with the intention of making the experiment less monotonous, and to stop participants from noticing the strict design features of the stimulus scenes. In particular, each participant saw: four scenes with twelve objects in all four colours, as opposed to the two-colour schemes; two scenes containing only three objects; and ten further filler scenes which intentionally violated the above design criteria. The filler scenes for each participant were chosen such that in eleven or twelve scenes the target was a cube instead of a ball, in two scenes the landmark was a ball, in four scenes there was no obvious landmark close to the target, in eight scenes the target was unique (i.e. it could not be described by its inherent visual properties alone), in nine or ten scenes the target and landmark shared type, and in two or three scenes target and landmark

were of the same size; for participants who saw Trial Set 2 all stimulus scenes also had a target and landmark of the same size.

The sequence of the 32 scenes that were shown to a particular participant was determined by the following three steps:

1. Pick the opposite trial set to the one that the last participant saw and randomise its order.

2. Pick the set of 16 filler scenes to be shown to this participant and randomise their order.

3. Interleave the two sets so that each stimulus scene is preceded by one filler scene.

After having described all 32 scenes in the trial, participants were asked to complete an exit questionnaire, which gave them the option of having their data discarded and asked for their opinion on whether the task became easier over time and any other comments they might wish to make.

The experiment was started by 318 native English speakers, of which 294 completed all 32 scenes. They were recruited by word of mouth via a widely-circulated call for participation and two electronic mailing lists.[3] The participants were predominantly in their twenties or thirties and mostly university-educated. A slight majority (54%) were female. None of them reported colour-blindness. Each referring expression in the corpus is tagged with an anonymous ID number linking it to some simple demographic data about the contributing participant, including gender, age, type of English spoken, and field of education.

[3]The Corpora List and the SIGGEN List.

16

## 4  Data Filtering and Annotation

Of the 294 participants who completed the experiment, five consistently used only type, although the target's type was never fully distinguishing in any of the stimulus scenes. For example, these participants described the target in Figure 1 simply as *ball*, which does not distinguish it from the two other balls in the scene. We discarded the data of these participants under the assumption that they had not understood the instruction that their descriptions were to uniquely identify the target. Two participants' data were discarded because they provided text that was unrelated to the displayed scenes. Of the remaining 287 participants, 140 saw Trial Set 2 and 147 saw Trial Set 1. The data from seven randomly-chosen participants from Trial Set 1 were discarded to balance the corpus in terms of the between-participant feature $TG\_Size = LM\_Size$. Each person described the 16 scenes contained in either of the trial sets, resulting in a corpus of 4480 descriptions in total, with 140 descriptions for each scene. No other corpus of referring expressions contains as many descriptions for each referential scenario from different speakers, which makes this corpus ideal for the study of speaker-specific preferences and non-deterministic choices in content selection.

Only five of the 4480 descriptions used the ternary spatial relation *between*, and one description mentioned two distinct spatial relations, one to the intended landmark and one to another object. The relation to the third object in these six descriptions was disregarded in the analysis presented here.

In order to be able to analyse the semantic content of the referring expressions, we semi-automatically annotated the inherent attributes and relations contained in each of them. The attributes annotated are

- type[ball, cube]
- colour[blue, green, red, yellow]
- size[large, small]
- location[right, left, front, top, bottom, centre]
- relation[horizontal, vertical]

Each attribute (except relation) is prefixed by either tg_ or lm_ to mark which of the objects it pertains to. For example, tg_size indicates that the size of the target was mentioned.

| attribute | count | % of total 4480 descriptions | % of all 600 relational descriptions |
|---|---|---|---|
| tg_size | 2587 | 57.8 | – |
| tg_colour | 4423 | 98.7 | – |
| tg_location | 81 | 1.8 | – |
| relation | 600 | 13.4 | – |
| lm_size | 327 | 7.3 | 54.5 |
| lm_colour | 521 | 11.6 | 86.8 |
| lm_location | 10 | 0.2 | 1.7 |

Table 1: **Attribute counts in GRE3D7**

In the 83 descriptions containing comparatives, such as Example (1), we ignored the second object that the target was being compared to. In all of these cases, the target's colour and type were also mentioned, which means that in the context of the simple scenes at stake here, Example (1) is semantically equivalent to Example (2).

(1)  the smaller of the two red balls

(2)  the small red ball

The question of how to deal with the relative nature of size is a separate, non-trivial, issue; see (van Deemter, 2000; van Deemter, 2006).

## 5  Analysis of the GRE3D7 Corpus

In this section we examine the content of the 4480 descriptions that make up the GRE3D7 Corpus. We first give an overview of the use of the non-relational attributes, and then proceed to investigate the hypotheses from Section 2 regarding the use of spatial relations.

The target object's type was mentioned in each description in the corpus, and each relational description contained the landmark object's type. Table 1 shows the number of descriptions containing each of the other attributes.

### 5.1  Sparing Use of location

Only 81 descriptions (1.8%) made reference to the target referent's location in the scene, as in Example (3); and of the 600 relational descriptions in the corpus, only ten (1.7%) contained the location of the landmark, as in Example (4).

(3)  the large yellow ball on the left [Scene 9]

(4) the small ball next to the large cube on the left hand side [Scene 6]

There were no descriptions containing both tg_location and lm_location. This might indicate that participants who used a relation were more likely to conceptualise the target–landmark pair as a unit with just one location rather than as two individual entities. However, the corpus was not designed to investigate this issue and the numbers for use of location are too low to draw any definite conclusions.

## 5.2 Abundant Use of colour

Colour was used in the vast majority of descriptions: 98.7% of all descriptions included the colour of the target object and 86.8% of the relational descriptions included the colour of the landmark object. A high number of descriptions containing colour could be expected, as colour was part of the shortest possible minimal description not containing any spatial information (we call this the *inherent* MD of the target) for 20 of the 32 scenes (all but Scenes 17–24 and 29–32). However, the fact that colour was also included in the majority of the descriptions containing spatial information, in the form of a relation or the location, confirms previous findings to the effect that colour is often included in descriptions redundantly (Belke and Meyer, 2002; Arts, 2004; Gatt, 2007).

## 5.3 Utilitarian Use of size

The target's size was mentioned in 57.8% of all descriptions, and the landmark's size in 54.8% of the relational descriptions.

Considering that tg_size was part of the inherent MD in only 12 of 32 scenes (37.5%) of the stimulus scenes (Scenes 2, 4, 9–12, 18, 20 and 25–28), 57.8% seems like a high proportion of descriptions to be using this attribute. The use of tg_size for scenes where it was part of the inherent MD was at 90.2% very high, but this only accounts for just under 60% of all the descriptions that contained this attribute. The remaining 40% of descriptions containing tg_size were given for scenes in which this attribute was not strictly necessary to distinguish the target from the other objects.

Findings from eye-tracking experiments in psycholinguistics have shown that size is rarely used in situations where it adds no discriminatory power to the referring expression at all, and that it is more likely to be used to compare to or distinguish from other objects of the same type, while the same is not true for colour (Sedivy, 2003; Brown-Schmidt and Tanenhaus, 2006). Let us therefore consider in particular the scenes where tg_size was not part of the inherent MD, and look at the differing utility of tg_size in these scenes: 12 of the 20 scenes where tg_size was not necessarily part of the inherent MD (Scenes 1, 3, 5–8, 13–16, 17, 19, 21–24 and 29–32) nonetheless contained another object that shared the target's type (ball) but not its size (Scenes 1, 3, 17, 19, 21–24 and 29–32). In these scenes, tg_size remains a useful attribute to use, even if tg_type is also included.

Based on the psycholinguistic findings mentioned above, one might expect that the use of tg_size is higher for these scenes because here it helps distinguish from another object of the same type rather than only from objects of a different type. This hypothesis is supported by the data: tg_size was used in 45.6% of the descriptions for scenes where it was not part of the inherent MD but there was another object of same type and different size as the target. For scenes where tg_size could only distinguish the target from objects of the other type, it was only used in 27.3% of cases ($\chi^2$=94.97, $df$=1, $p \ll .01$).

## 5.4 The Use of Spatial Relations

600 of the 4480 descriptions in the GRE3D7 Corpus (13.4%) mentioned a spatial relation. This was despite the fact that spatial information was not required in any of the stimulus scenes. Most existing approaches to spatial relations in REG would therefore never include a relation for any of the stimuli.

In this section, we examine the circumstances under which the participants of the GRE3D7 data collection experiment used the spatial relation between the target object and the intended landmark. We will first examine participant-dependent and temporal factors and then move on to analyse the impact that the design features of the scenes, described in Section 2, had on the use of relations.

### General Factors

We first checked for broad participant-dependent preferences for or against using relations in the

GRE3D7 Corpus. The behaviour of participants who use an exclusive strategy of either always or never including a relation in their referring expressions would be easy to predict in a computational model and does not contribute to any variation across different scenes. In order to gain a clear understanding of this variation, we will concentrate on the data from participants who varied their use of relations between scenes.

Half of the participants (50.3%) adopted an exclusive strategy regarding the use of relations. However, the split between the two exclusive strategies was very uneven: 135 participants never used a spatial relation and only six used a spatial relation for all 16 stimulus scenes they saw. In the following, we analyse the data from the 139 participants who used a relation for some scenes but not for others. On average, these participants used a relation in 22.7% of their descriptions.

In (Viethen and Dale, 2008), we observed a 'laziness effect' whereby participants' use of relations decreased over the course of the experiment. A number of participants mentioned in the exit interview that they noticed over time that relations were never required and stopped using them. Such a conscious, or semi-conscious, adjustment masks people's natural propensity to use a relation in a reference situation where they come anew at the task rather than describing one object after another.

In the GRE3D7 collection experiment, each participant saw eight filler scenes in which spatial relations were required to distinguish the target. These filler scenes were included to stop participants from consciously noticing that relations were never required in the stimulus scenes. We hoped that this would reduce the laziness effect and thereby produce results that better approximate people's natural tendency to use a relation. However, Figure 3 shows that, despite the use of these filler scenes, the use of relations declined over the course of the experiment. Participants who did not follow an exclusive strategy clearly used more relations for scenes they saw early on than for those they saw towards the end. We divided the data set into quartiles in order to test the statistical significance of this decline. The falling trend was statistically significant at $p \ll .01$ ($\chi^2 = 55.42$, $df = 3$). However, any temporal effect in GRE3D7 should not interfere with
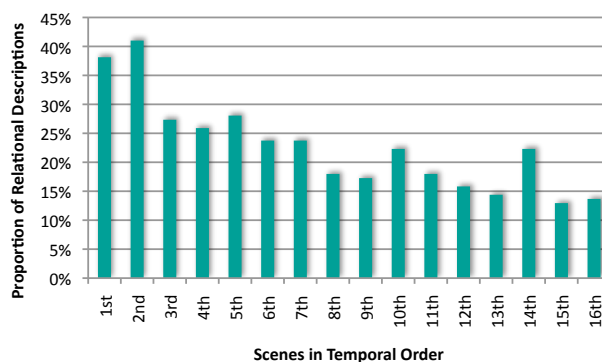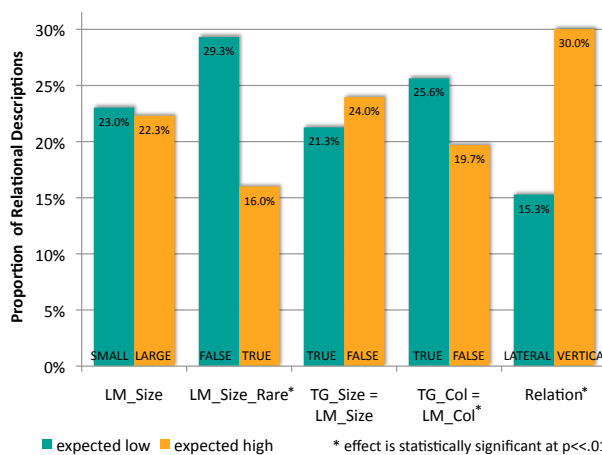


Figure 3: **Temporal effect on use of** relation



Figure 4: **Effect of design variables on use of** relation

between-stimulus effects, as the stimuli were presented in a randomised order.

### Influence of Scene Features on Relation Use

We will now turn to the examination of **Hypotheses 1–5** from Section 2. Figure 4 shows the impact that each of the five variables of the scene design had on the use of relations. The left (green) columns represent the conditions for which we expected fewer relations to be used, and the right (yellow) columns represent the conditions for which we expected a higher use of relations, according to **Hypotheses 1–3** and **5**. **Hypothesis 4** expected the reverse results for TG_Size = LM_Size and TG_Col = LM_Col. All factors except LM_Size and TG_Size = LM_Size had a statistically significant effect.

**Hypotheses 1** and **2**, which expected a large landmark with a rare or unique size to be more salient and therefore more likely to be used, are not supported by the data here. LM_Size did not have a reliable effect ($\chi^2 = 0.16$, $df = 1$, $p > .6$) and

19

LM_Size_Rare shows the opposite effect of the one we expected: a relation to a landmark with a common size is significantly more likely to be included in a referring expression than one to a landmark with a rare or unique size ($\chi^2$=56.19, $df$=1, $p \ll .01$). On closer inspection, this is likely to be due to a factor that was not explicitly tested or controlled for in this experiment: the length of the inherent MD of the target referent. In most scenes with a common landmark size (all but Scenes 1, 3, 17, and 19), all three inherent attributes (size, colour and type) are necessary to distinguish the target from the other objects without using locational information. In all scenes where the landmark's size is rare or unique, colour and type suffice. In other words, targets which are harder to describe using inherent visual properties only are more likely to be described by a relation to a nearby landmark.

**Hypotheses 3** and **4** predicted two mutually exclusive scenarios based on the assumption that the similarity between the target and the landmark object is of special importance, as the participant's visual attention is likely to be focussed on these two objects. **Hypothesis 3** predicted that a visual difference between the landmark and the target would increase the landmark's salience and therefore the use of the spatial relation to this landmark. **Hypothesis 4** predicted that high visual similarity between target and landmark might result in these two objects being conceptualised as a unit, which would increase the likelihood of both objects being mentioned. The target and landmark object were always of different types, so their similarity depends on their size and their colour, captured in the variables TG_Size = LM_Size and TG_Col = LM_Col. TG_Size = LM_Size did not show a significant effect on the use of relations ($\chi^2$=2.29, $df$=1, $p > .1$). The effect of TG_Col = LM_Col favours **Hypothesis 4**, as a landmark of the same colour as the target is more likely to be included in the target's description than one that has a different colour from the target ($\chi^2$=11.18, $df$=1, $p \ll 0.01$).

The variable Relation had the expected effect: A vertical relation is significantly more likely to be used than a lateral one ($\chi^2$=69.00, $df$=1, $p \ll .01$). This confirms **Hypotheses 5**.

# 6   Conclusion

We have described the GRE3D7 Corpus, a collection of human-produced distinguishing descriptions that is considerably larger than any other existing corpus. The collection also uses scenes that are a degree more complex than those found in existing corpora; these are based on a principled design in order to provide a measure of control over what can be learned from the data. In this paper we have described the details of the collection experiment and have presented an analysis of the impacts that the design variables had on the content of the resulting descriptions. The main outcomes of this analysis are:

*Colour is used in 99% of all descriptions.* It is also used redundantly in 87% of all relational descriptions. This is in accordance with findings in other corpora and psycholinguistic studies.

*Size is used when it is distinguishing.* The size of the target referent was much more likely to be included when it was useful in distinguishing from another object in the scene, especially those of the same type.

*Just over half of the participants follow an exclusive strategy for the use of relations.* A large proportion of participants (135) opted to never use a relation, while a much smaller number of people (6) used a relation in all of their descriptions. The remaining 139 participants are responsible for the variation in the data, as they used a relation to describe the target in some but not all scenes.

*The target–landmark relation is used more often if it is vertical than if it is lateral.* This confirms previous psycholinguistic findings showing that humans prefer vertical relations and prepositions over horizontal, and in particular lateral, ones.

*If a landmark shares colour with the target it is more likely to be used in a referring expression.* This lends support to the hypothesis that visual similarity between target and landmark increases the likelihood of the relation between them being used.

The data thus sheds additional light on the nature of human-produced descriptions of objects in visual scenes. It also, of course, provides a rich corpus of data that can be readily used to evaluate the performance of computational algorithms for the generation of referring expressions.

20

# References

Anja Arts. 2004. *Overspecification in Instructive Texts*. Ph.D. thesis, University of Tilburg, The Netherlands.

Eva Belke and Antje S. Meyer. 2002. Tracking the time course of multidimensional stimulus discrimination: Analysis of viewing patterns and processing time during same-different decisions. *European Journal of Cognitive Psychology*, 14(2):237–266.

Anja Belz and Albert Gatt. 2007. The Attribute Selection for GRE Challenge: Overview and evaluation results. In *Proceedings of the Workshop on Using Corpora for NLG: Language Generation and Machine Translation (UCNLG+MT)*, pages 75–83, Copenhagen, Denmark.

Sarah Brown-Schmidt and Michael K. Tanenhaus. 2006. Watching the eyes when talking about size: An investigation of message formulation and utterance planning. *Journal of Memory and Language*, 54:592–609.

David J. Bryant, Barbara Tversky, and Nancy Franklin. 1992. Internal and external spatial frameworks representing described scenes. *Journal of Memory and Language*, 31:74–98.

David J. Bryant, Barbara Tversky, and M. Lanca. 2000. Retrieving spatial relations from observation and memory. In Emile van der Zee and Urpo Nikanne, editors, *Cognitive interfaces: Constraints on linking cognitive information*, pages 94–115. Oxford University Press, Oxford, UK.

Robert Dale and Nicholas Haddock. 1991. Content determination in the generation of referring expressions. *Computational Intelligence*, 7(4):252–265.

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, Vancouver BC, Canada.

Klaus-Peter Gapp. 1995. Angle, distance, shape, and their relationship to projective relations. In *Proceedings of the 17th Annual Meeting of the Cognitive Science Society*, pages 112–117, Pittsburgh PA, USA.

Claire Gardent, Hélène Manuélian, Kristina Striegnitz, and Marilisa Amoia. 2004. Generating definite descriptions: Non incrementality, inference and data. In Thomas Pechmann and Christopher Habel, editors, *Multidisciplinary Approaches to Language Production*, pages 53–86. Walter de Gruyter, Berlin, Germany.

Claire Gardent. 2002. Generating minimal definite descriptions. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 96–103, Philadelphia PA, USA.

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 11th European Workshop on Natural Language Generation*, pages 49–56, Schloß Dagstuhl, Germany.

Albert Gatt, Anja Belz, and Eric Kow. 2008. The TUNA Challenge 2008: Overview and evaluation results. In *Proceedings of the 5th International Conference on Natural Language Generation*, pages 198–206, Salt Fork OH, USA.

Albert Gatt, Anja Belz, and Eric Kow. 2009. The TUNA-REG Challenge 2009: Overview and evaluation results. In *Proceedings of the 12th European Workshop on Natural Language Generation*, pages 174–182, Athens, Greece.

Albert Gatt. 2007. *Generating Coherent Reference to Multiple Entities*. Ph.D. thesis, University of Aberdeen, UK.

Surabhi Gupta and Amanda Stent. 2005. Automatic evaluation of referring expression generation using corpora. In *Proceedings of the Workshop on Using Corpora for Natural Language Generation*, pages 1–6, Brighton, UK.

Helmut Horacek. 2003. A best-first search algorithm for generating referring expressions. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, pages 103–106, Budapest, Hungary.

Helmut Horacek. 2004. On referring to sets of objects naturally. In *Proceedings of the 3rd International Conference on Natural Language Generation*, pages 70–79, Brockenhurst, UK.

Pamela W. Jordan and Marilyn Walker. 2005. Learning content selection rules for generating object descriptions in dialogue. *Journal of Artificial Intelligence Research*, 24:157–194.

John Kelleher and Geert-Jan Kruijff. 2006. Incremental generation of spatial referring expressions in situated dialog. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pages 1041–1048, Sydney, Australia.

Emiel Krahmer and Mariët 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, Sebastiaan van Erk, and André Verleg. 2003. Graph-based generation of referring expressions. *Computational Lingustics*, 29(1):53–72.

Barbara Landau. 2003. Axes and direction in spatial language and spatial cognition. In Emilie van der Zee

and Jon M. Slack, editors, *Representing Direction in Language and Space*, pages 18–38. Oxford University Press, Oxford, UK.

John Lyons. 1977. *Semantics*, volume 2. Cambridge University Press, Cambridge, UK.

Julie C. Sedivy. 2003. Pragmatic versus form-based accounts of referential contrast: Evidence for effects of informativity expectations. *Journal of Psycholinguistic Research*, 32(1):3–23.

Thora Tenbrink. 2004. Identifying objects on the basis of spatial contrast: An empirical study. In Christian Freksa, Markus Knauff, Bernd Krieg-Brckner, Bernhard Nebel, and Thomas Barkowsky, editors, *Spatial cognition IV: Reasoning, action, interaction*, number 3343 in Lecture Notes in Computer Science, pages 124–146. Springer, Berlin/Heidelberg, Germany.

Kees van Deemter. 2000. Generating vague descriptions. In *Proceedings of the 1st International Conference on Natural Language Generation*, pages 179–185, Mitzpe Ramon, Israel.

Kees van Deemter. 2006. Generating referring expressions that involve gradable properties. *Computational Linguistics*, 32(2):195–222.

Ielka van der Sluis and Emiel Krahmer. 2004. Evaluating multimodal NLG using production experiments. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, Lisbon, Portugal, 26-28 May.

Ielka van der Sluis and Emiel Krahmer. 2005. Towards the generation of overspecified multimodal referring expressions. In *Proceedings of the Symposium on Dialogue Modelling and Generation at the 15th Annual Meeting of the Society for Text and Discourse*, Amsterdam, The Netherlands, 6-9 July.

Ielka van der Sluis. 2005. *Multimodal Reference, Studies in Automatic Generation of Multimodal Referring Expressions*. Ph.D. thesis, Tilburg University, The Netherlands.

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.

Jette Viethen, Simon Zwarts, Robert Dale, and Markus Guhe. 2010. Dialogue reference in a visual domain. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, Valetta, Malta.

Steven Yantis. 1998. Control of visual attention. In Harold Pashler, editor, *Attention*, chapter 6, pages 223–256. Psychology Press, Hove, UK.