COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 37 (2013) 395–411

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/cogs.12019

The Effect of Scene Variation on the Redundant Use of Color in Definite Reference

Ruud Koolen, Martijn Goudbeek, Emiel Krahmer

Tilburg Center for Cognition and Communication (TiCC), Tilburg University

Received 13 October 2011; received in revised form 25 April 2012; accepted 16 May 2012

Abstract

This study investigates to what extent the amount of variation in a visual scene causes speakers to mention the attribute color in their definite target descriptions, focusing on scenes in which this attribute is not needed for identification of the target. The results of our three experiments show that speakers are more likely to redundantly include a color attribute when the scene variation is high as compared with when this variation is low (even if this leads to overspecified descriptions). We argue that these findings are problematic for existing algorithms that aim to automatically generate psychologically realistic target descriptions, such as the Incremental Algorithm, as these algorithms make use of a fixed preference order per domain and do not take visual scene variation into account.

Keywords: Definite reference; Overspecification; Scene variation; Computational models

1. Introduction

In everyday language use, speakers often produce definite descriptions of *target* objects (such as "the brown chair"), and they aim to do this in such a way that their addressee is able to uniquely distinguish the target from its surrounding *distractor* objects (e.g., other furniture items). It is well known that speakers tend to *overspecify* their descriptions by providing *redundant* attributes that are not needed for target identification (e.g., Arts, 2004; Arts, Maes, Jansen, & Noordman, 2011; Engelhardt, Bailey, & Ferreira, 2006; Koolen, Gatt, Goudbeek, & Krahmer, 2011), which is, as some have argued, in conflict with the Maxim of Quantity proposed by Grice (1975). In any case, overspecification is hard to capture for Referring Expression Generation (REG) algorithms, which are computational models that aim to generate definite object descriptions (Reiter & Dale, 2000).

Correspondence should be sent to Ruud Koolen, Tilburg Center for Cognition and Communication (TiCC), Faculty of Humanities, Tilburg University, P.O. Box 90153, NL-5000 LE Tilburg, The Netherlands. E-mail: R.M.F.Koolen@uvt.nl

These algorithms generally focus on *Content Determination*: what attributes should be included to distinguish the target? Most algorithms (including Dale and Reiter (1995) influential Incremental Algorithm, introduced in this journal) rest on the assumption that some attributes are preferred over others, and that these preferences are fixed for a given domain. As we explain later on, this assumption gives rise to some amount of overspecification. However, given that there is growing awareness that speech production and visual scene perception are closely intertwined (e.g., Griffin & Bock, 2000; Hanna & Brennan, 2007; Meyer, Sleiderink, & Levelt, 1998), we argue that whether speakers redundantly mention a certain attribute (in particular, color) does not merely depend on how preferred this attribute is in a given domain but also on the visual variation in a scene.

1.1. Psycholinguistic evidence for overspecification

It is often assumed that speakers tend to obey the Maxim of Quantity (Grice, 1975), stating that speakers should make their contribution as informative as is required (for the current purpose of the exchange), but not more informative than that. If one regards identification as a core purpose of uttering a target description (as we do here, following many previous articles on overspecification, e.g., Engelhardt et al., 2006; Olson, 1970; Pechmann, 1989), the Maxim of Quantity would result in descriptions that contain just enough attributes for the addressee to identify the target. This prediction is at odds with the observation that speakers often overspecify and provide their addressees with redundant attributes (e.g., Arts, 2004, 2011; Engelhardt et al., 2006; Koolen et al., 2011). One plausible reason for this is that speakers tend to include attributes that they prefer, even if mentioning these attributes leads to overspecified target descriptions. Attributes that are known to be generally preferred tend to be perceptually salient, for example, color (e.g., Belke & Meyer, 2002; Pechmann, 1989). The focus in this study lies therefore on the redundant use of the attribute color in definite object descriptions.

1.2. Overspecification in Referring Expression Generation

The above findings concerning the occurrence of referential overspecification, and speakers' preferences for certain attributes have important implications for researchers in the field of Natural Language Generation (NLG). NLG is a subfield of Artificial Intelligence that aims to build models for automatically generating natural language text or speech, usually from nonlinguistic information (e.g., from a database; Reiter & Dale, 2000). NLG systems typically use REG algorithms that generate distinguishing descriptions of objects (Mellish et al., 2006). Many of these REG algorithms can be seen as (implicit or explicit) computational interpretations of the Gricean maxims of conversational implicature. For example, the algorithms discussed by Dale and Reiter (1995) explicitly take the Maxim of Quantity as a starting point, aiming to approximate the referential behavior of human speakers.

So, how should the Maxim of Quantity be interpreted in the context of an REG task? It is worth mentioning in this respect that most current REG algorithms generate referring

expressions that are solely intended to identify a target object and have no other (non-identificational) communicative purposes (such as giving a warning). With that in mind, Dale and Reiter (1995, p. 240) propose the following interpretation of the Maxim of Quantity: "a referring expression should contain enough information to enable the hearer to identify the object referred to, but not more information." Many of the various REG algorithms that have been proposed so far rely on this interpretation. For example, the *Full Brevity* Algorithm (Dale, 1989, 1992) is based on a strict interpretation of the Maxim of Quantity and always seeks to find the shortest possible target description (in terms of the number of attributes included). This is not necessarily the case with the *Greedy Heuristic* algorithm (Dale, 1989, 1992), which iteratively selects the attribute that rules out most of the distractor objects at each stage of the attribute selection process. However, the most influential REG algorithm to date (as discussed by Van Deemter, Gatt, Van der Sluis, & Power, 2012) is arguably the *Incremental Algorithm* (IA). In this study, we therefore base our predictions mainly on this algorithm.

The IA (Dale & Reiter, 1995) generates target descriptions by using a predetermined *preference order*. This is a ranking of all attributes that can possibly occur in a given domain, where preferred attributes are ranked before less-preferred attributes. Dale and Reiter (1995) argue that this preference order is fixed for every domain, and that it can typically be determined empirically. To illustrate how a preference order is usually determined, let us consider the *furniture domain*, which has often been used before in REG studies (e.g., Gatt & Belz, 2010; Gatt, Van der Sluis, & Van Deemter, 2007; Van Deemter, Gatt, Van der Sluis, et al., 2012), and which we also use in this study. Empirical data from (among others) Koolen et al. (2011) show that the target's type (e.g., "chair") is practically always mentioned. Therefore, type can be placed at the head of the preference order. The second most frequent attribute in this domain is color, while infrequent attributes such as size and orientation occur in the tail of the preference order.

How does the IA use this preference order? Consider the two chairs depicted in Fig. 1, and imagine that the algorithm wants to distinguish the brown chair from the green chair.

Fig. 1. A brown chair and a green chair. Manipulations of color may not be visible in a black and white print of this paper.

In this case, the IA would first consider the type attribute, because it is at the head of the preference order. As type does not rule out the only distractor (because both objects are chairs), the IA considers the next attribute in line (color) and checks whether the attribute–value pair <color, brown> rules out the distractor, which it does. The resulting description could be realized as "the brown one". However, as most descriptions tend to contain a head noun mentioning a target's type, the IA always adds the type to a description (note that it only does this when type was not selected at an earlier stage, like in our example).

The IA does not backtrack for overspecification, which means that it does not remove selected attributes that turn out to be redundant in the end. This is the case when there exists one other—less preferred—attribute (or a combination of several other attributes) that renders all higher ranked attributes obsolete. For example, one can think of a visual scene with one target object and two distractors, where color rules out one distractor and size two. Color and size are then both selected, although including size would have been sufficient. In this way, the IA is able to generate overspecified descriptions.

1.3. The current study

To what extent do algorithms like the IA produce descriptions that are psychologically realistic? This question is often raised in the field of REG, particularly when the output of algorithms is evaluated against human corpus data (e.g., Gatt & Belz, 2010). In this study, we compare human object descriptions with those of REG algorithms like the IA, where we focus on overspecification: To what extent are automatically generated descriptions comparable to human descriptions in terms of the redundant attributes that they contain?

We expect to find at least one important difference between automatically generated and human target descriptions, and we expect this difference to be related to the visual variation that is present in a scene. There is growing awareness that visual scene perception and speech production are closely related (e.g., Griffin & Bock, 2000; Hanna & Brennan, 2007; Meyer et al., 1998), but little is known about how scene perception influences the production of referring expressions, and, typically, existing REG algorithms such as the IA pay no attention to visual information in a scene. We hypothesize that human speakers are sensitive to *scene variation*, which we operationalize as the number of dimensions in which the objects in a scene differ. For example, reconsider the furniture domain, in which there might be scenes where objects differ in terms of only one dimension (e.g., type), but also scenes in which objects differ in terms of several dimensions (e.g., type, color, orientation, and size). Arguably, describing a target in the latter case is a more difficult task and might therefore cause speakers to include more redundant attributes in their target descriptions (resulting in more overspecification). On the basis of existing research discussed earlier, we expect this to be at least the case for the preferred attribute color.

The IA does not necessarily include more redundant attributes when the objects in a scene differ across more dimensions. One situation in which this problem becomes

apparent is when type is sufficient to distinguish the target: given that, in our domain, type is at the head of the preference order and hence selected first, the IA would not select a (redundant) color attribute, irrespective of the visual variation in a scene.

To investigate whether human speakers include color more often in high-variation scenes, we performed three experiments in which participants were presented with visual scenes consisting of eight objects including one target, asking them to produce distinguishing descriptions for the target objects. The experiments had two conditions (high and low visual variation), and we made sure that color was never needed for target identification. We investigate whether human speakers use color more frequently in the high-variation condition and conclude with contrasting our findings with the predictions of the IA.

2. Experiment 1

2.1. Method

2.1.1. Participants

Participants were 42 undergraduate students who participated in pairs. Twenty-one students (11 female, mean age = 21 years and 7 months) acted as speakers, the other 21 as addressees. All participants were native speakers of Dutch (the language of the study) and participated for course credits.

2.1.2. Materials

The stimulus material consisted of artificially constructed pictures of furniture items,[1] which have been used before extensively in the field of REG (e.g., Van Deemter, Gatt, Van der Sluis, et al., 2012). The furniture items could vary in terms of four attributes and their corresponding values. All possible attribute–value pairs are listed in Table 1.

The critical trials all contained eight furniture items: one target object and seven distractor objects. The target objects were clearly marked by black borders so that the speakers could easily distinguish them from the distractor objects. The furniture items were randomly positioned on a computer screen in a two (row) by four (column) picture grid.

Experiment 1 had two conditions. The critical trials in the *low-variation condition* were constructed in such a way that there was limited variation between the target and

Table 1
Attributes and possible values of the furniture items

Attributes	Possible Values
Type	Chair, sofa, fan, television, desk
Color	Red, blue, green, brown, gray
Orientation	Front, back, left, right
Size	Large, small

the distractor objects: The furniture items differed only in terms of the attribute type. In the *high-variation condition,* the target and the distractor objects differed in terms of all four possible attributes (i.e., type, color, size, and orientation). Mentioning type was sufficient to successfully distinguish the target in all critical trials in the two conditions, which implies that including color was never needed to distinguish the target. Note also that the IA would not include color in either of the two conditions: as including type (which is at the head of the preference order in this domain) is sufficient for distinguishing the target, the IA would not include any further attributes in line (such as color). Fig. 2 depicts examples of critical trials in the two respective conditions.

There were 20 critical trials (10 per condition) and 40 fillers. We made one block of 60 trials in a fixed random order (which was presented to one half of the speakers), and a second block containing the same trials in reverse order (which was presented to the other half of the speakers). The fillers consisted of four pictures of Greebles (Gauthier & Tarr, 1997): one clearly marked target referent and three distractor objects, all positioned in a two by two picture grid. Greebles are complex and difficult to refer to, which made them useful fillers in our experiment. The Greebles could not be distinguished in terms of their color because they were all in the same color every time (so speakers were not primed with the attribute color when describing the fillers).

2.1.3. Procedure

The experiment was performed in an experimental laboratory. After the two participants had arrived in the room, it was randomly decided who was going to act as the speaker and who as the addressee. Thereafter, the participants were seated opposite to each other. The speaker was presented with the 60 trials on a computer screen and was asked to describe the target referents in such a way that the addressee would be able to uniquely identify them. There were two practice trials. The instructions emphasized that it would not make sense to include location information in the descriptions, as the addressee was presented with the pictures in a different order. The speaker could take as much time as needed to describe the target, and his or her target descriptions were recorded

Fig. 2. Examples of critical trials in Experiment 1: for the low-variation condition (left picture) and for the high-variation condition (right picture). Manipulations of color may not be visible in a black and white print of this paper.

with a voice recorder. The addressee was presented with the same 60 trials as the speaker in a paper booklet and was asked to mark the picture that he or she thought the speaker was describing on an answering form. The instructions emphasized that the addressee was—to a limited extent—allowed to ask for clarification: It was allowed to ask the speaker to give more information or to repeat information that had already been given, but not to ask for specific information (i.e., specific attributes). Because of the small number of clarification requests and thus clarifications by our speakers (asking for clarification occurred in only 1.1% of the critical trials), the data presented here should be regarded as initial reference. Once the addressee had identified a target, this was communicated to the speaker, who then went on to describe the next one. After completion of the experiment, none of the participants indicated that they had been aware of the actual goal of the study. All found it an easy task to accomplish.

2.1.4. Design and statistical analysis

Experiment 1 had a within-participants design with scene variation (levels: low, high) as the independent variable, and the proportion of descriptions containing a color attribute as the dependent variable. As described above, we made sure that speakers never needed to include color in their target descriptions to produce a distinguishing description of the target. Thus, if speakers did mention color anyway, this caused the expression to be overspecified.

Our statistical procedure consisted of two repeated-measures ANOVAS: one on the participant means (F_1) and one on the item means (F_2).

2.2. Results

In total, 420 target descriptions were produced in this experiment. All of these contained a type attribute and were fully distinguishing. Fig. 3 depicts the proportion of

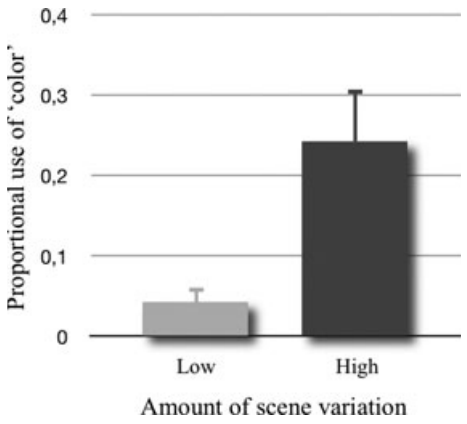

Fig. 3. Results for Experiment 1: the proportion of referring expressions (plus standard deviations) containing a "color" attribute as a function of the variation in the visual scene.

expressions that contained a color attribute as a function of the condition in which the descriptions were uttered.

As hypothesized, the scene variation affected the proportional use of the redundant attribute color (F_1 (1,20) = 12.537, $p < .01$; F_2 (1,18) = 23.416, $p < .001$). More specifically, speakers were more likely to include color when there was high variation in the picture grid ($M = 0.24$, $SD = 0.07$) as compared with when this variation was low ($M = 0.04$, $SD = 0.02$).

Experiment 1 confirmed our hypothesis about the role of scene variation on speakers' tendencies to redundantly include a color attribute in their target descriptions. In the next experiment, we will see whether the same applies when the difference between the low- and high-variation conditions becomes more subtle (in terms of amount of variation).

3. Experiment 2

3.1. Method

3.1.1. Participants

Participants were again Dutch-speaking undergraduate students who participated in pairs. This time, there were 22 students who acted as speakers (12 female, mean age = 22 years and 4 months). None of these speakers had acted as a speaker in Experiment 1. Another 22 students acted as addressees in this experiment. Most of these had been speakers in Experiments 1 or 3; in a few cases, the addressee was a confederate.

3.1.2. Materials

There were 20 critical trials in two conditions, and these trials all contained one clearly marked target referent and seven distractor objects. We used the same fillers as before. Again, there was maximum variation between the target and the distractor objects in the *high-variation condition* (thus, the objects again differed in terms of the attributes type, color, orientation, and size). However, unlike in Experiment 1 (where the objects only had different types), the pictures in the *low-variation condition* now varied in terms of three attributes (type, orientation, and size) instead of one. This caused the difference between the trials in the two conditions to be more subtle than in Experiment 1. Fig. 4 depicts examples of trials in the two conditions of Experiment 2.

In all critical trials, mentioning type plus one other attribute (orientation or size but never color) was sufficient to produce a distinguishing description of the target. For example, in Fig. 4, a speaker could distinguish the target objects in both conditions by mentioning type and size. There were 10 trials in each condition, with an equal number of size and orientation trials, and again, mentioning color was never needed to distinguish the target. As in Experiment 1, the trials were built in such a way that the IA would not include color in its descriptions. In the low-variation condition in Fig. 4, the IA would not select color because all objects have the same color. In the high-variation condition in Fig. 4, the IA would first select type, because it is at the head of the preference order

Fig. 4. Examples of critical trials in Experiment 2: for the low-variation condition (left picture) and for the high-variation condition (right picture). Manipulations of color may not be visible in a black and white print of this paper.

in this domain. As both remaining distractors are then brown chairs, the algorithm will not include color and select size instead.

3.1.3. Procedure, design, and statistical analysis

As above.

3.2. Results

In total, 440 target descriptions were produced in this experiment. All of these contained a type attribute, and most (99.5%) were fully distinguishing. Fig. 5 depicts the proportion of expressions that contained a color attribute as a function of the condition in which the descriptions were uttered.

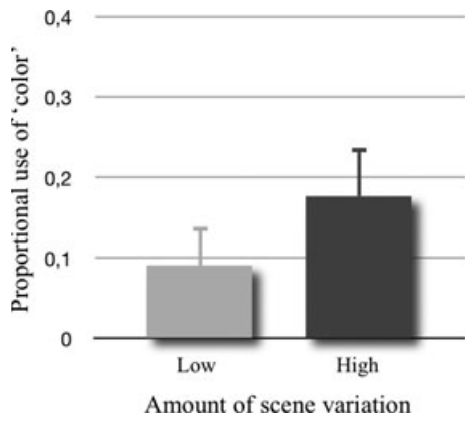

Fig. 5. Results for Experiment 2: the proportion of referring expressions (plus standard deviations) containing a "color" attribute as a function of the variation in the visual scene.

The general picture of the results of this experiment is comparable to that of the results of Experiment 1. We again found that the amount of variation between the target and the distractors affected the number of times that speakers included the redundant attribute color in their referring expressions (F_1 (1,21) = 7.092, $p < .05$; F_2 (1,18) = 10.515, $p < .01$). More specifically, the results showed that speakers were more likely to mention color when the variation in the picture grid was high ($M = 0.18$, $SD = 0.06$) as compared to when it was low ($M = 0.09$, $SD = 0.05$).

The results of Experiment 2, like those of Experiment 1, confirmed our hypothesis and indicate that speakers more often redundantly include color in their descriptions when the variation in a visual scene is high. In the next experiment, we will take a closer look at the role of the attribute type: Are speakers more likely to redundantly include color when the objects in a scene have different types as compared to when all objects are of the same type?

4. Experiment 3

4.1. Method

4.1.1. Participants

Participants were Dutch-speaking undergraduate students, again participating in pairs. There were 20 participants who acted as speakers (14 female, mean age = 22 years and 2 months). None of these participants had acted as a speaker in Experiments 1 or 2. Another 20 students acted as addressees, most of whom had been speakers in Experiments 1 or 2. In a few cases the addressee was a confederate.

4.1.2. Materials

Critical trials and fillers were constructed as above. Here, the crucial manipulation was that all scene objects had the same type in the *low-variation condition* (e.g., eight chairs), while they had different types in the *high-variation condition*. Furthermore, in the low variation condition, the objects varied in terms of their orientation and size, while in the high variation condition, they had different types, colors, orientations, and sizes. Fig. 6 depicts examples of trials in the two respective conditions.

In all critical trials, mentioning type plus one additional attribute (orientation or size but never color) was sufficient to uniquely distinguish the target. For example, in Fig. 6, a speaker could distinguish the targets in both conditions by mentioning type and orientation. There were 10 trials in each condition, with an equal number of size and orientation trials, and again, color was never needed to produce a distinguishing description. Like in the previous experiments, the critical trials were constructed in such a way that the IA would not include color in its descriptions. In the low variation condition (left picture in Fig. 6), the algorithm would skip both type and color (since they do not exclude any of the distractors), would then select orientation, and finally would add type to make the description a proper noun phrase. In the high variation condition (right picture in Fig. 6),

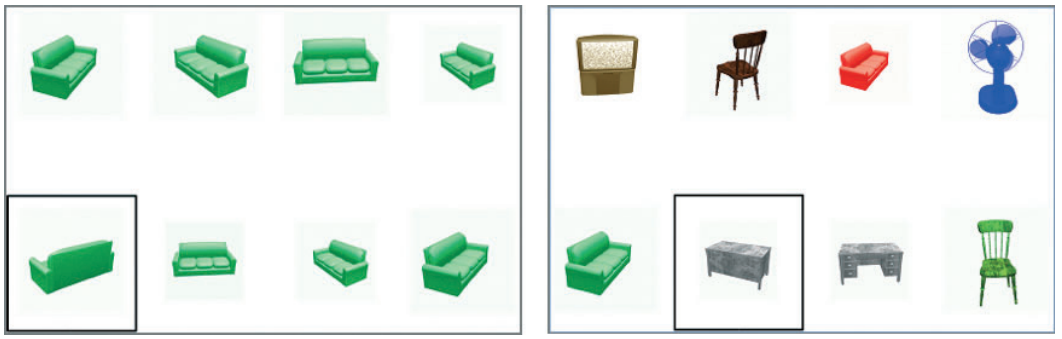

Fig. 6. Examples of critical trials in Experiment 3: for the low-variation condition (left picture) and for the high-variation condition (right picture). Manipulations of color may not be visible in a black and white print of this paper.

the IA would first select type. Since both remaining objects in this example are now desks of the same color and size, the algorithm would select orientation instead of color.

4.1.3. Procedure, design, and statistical analysis

As above.

4.2. Results

In total, 400 target descriptions were produced in the current experiment. Most of these contained a type attribute (94.3%), and the vast majority (97.5%) was fully distinguishing. Fig. 7 depicts the proportion of referring expressions that contained a color attribute as a function of the condition in which the expressions were uttered.

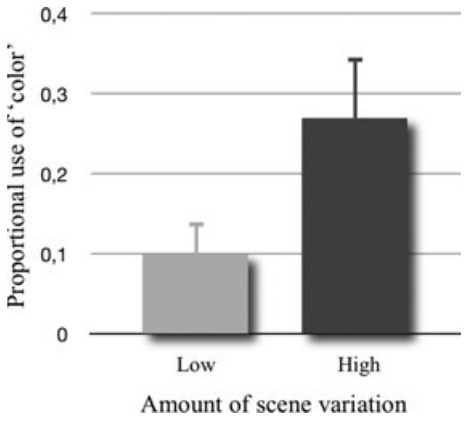

Fig. 7. Results for Experiment 3: the proportion of referring expressions (plus standard deviations) containing a "color" attribute as a function of the variation in the visual scene.

The results of Experiment 3 show a similar pattern as those of Experiments 1 and 2. We again found that the amount of variation in a visual scene affected the number of times that speakers redundantly included color in their target descriptions (F_1 (1,19) = 7.616, $p < .05$; F_2 (1,18) = 20.643, $p < .001$). Speakers were more likely to mention color when the variation in the scene was high ($M = 0.27$, $SD = 0.08$) than when it was low ($M = 0.10$, $SD = 0.04$). These results are again in line with our hypothesis that a higher amount of scene variation causes speakers to more frequently mention color redundantly in their descriptions.

5. Meta-analysis and speaker variation

5.1. Meta-analysis

The results of the three experiments reported so far all confirm our hypothesis that human speakers more often redundantly use color when presented with high-variation scenes than with low-variation scenes. The low-variation scenes of our three experiments tested scene variation in different, but related ways: In Experiment 1, the objects differed only in terms of their types; in Experiment 2, there was variation in terms of type, orientation, and size; and in Experiment 3, the objects were all of the same type. Did these differences lead to different use of color?

As we have seen, the proportion of referring expressions that contained a color attribute in low-variation scenes was slightly lower in the first experiment ($M = 0.04$) as compared with the second ($M = 0.09$) and third experiments ($M = 0.10$). We ran a statistical meta-analysis to find out whether these proportions were significantly different, combining data of the three experiments. To do this, we performed a 3 × 2 repeated-measures ANOVA (using Tukey's HSD for multiple comparisons), with Experiment as a between-subjects variable (levels: Experiments 1, 2, and 3) and Condition as a within-subjects variable (levels: low and high variation). As expected, Condition had a significant influence on the redundant use of color (F (1,60) = 26.727, $p < .001$). However, interestingly, we neither found a main effect of Experiment (F (2,60) = .302, *ns*) nor an interaction between Experiment and Condition (F (2,60) = 1.373, *ns*). This suggests that the effects reported in this article generalize over the different manipulations of variation in the low-variation conditions.

5.2. Speaker variation

Although the results of all three experiments show a similar pattern regarding the effects of condition, it might be that there are more individual differences between speakers in one experiment as compared with another. Arguably, individual differences between speakers are an interesting challenge for researchers in the field of REG (Dale & Viethen, 2010). Therefore, to see whether there was variation between the speakers in our three experiments, we drew scatter plots showing redundant color use of all speakers.

More specifically, we calculated the difference between the proportional use of color in the high- and the low-variation conditions for each speaker. For example, if a speaker mentioned color once when presented with low-variation scenes, and five times when presented with high-variation scenes, this person scored 4 in terms of the difference between the two conditions. As there were 10 trials in each condition, the scores for each speaker could range from −10 to 10. Fig. 8 depicts the individual scores for all participants that took part in our experiments.

Fig. 8 shows that the vast majority of the speakers scored between 0 and 5, meaning that these speakers mentioned more (or as much) redundant color attributes in the high-variation condition as compared with the low-variation condition. Therefore, we can be fairly certain that the effects reported in this article are consistent across participants, and that they were not driven by one or two individuals.

6. General discussion

We have shown that the amount of visual variation in a scene affects the number of times that speakers redundantly include color in their descriptions. In all three experiments, participants mentioned the color of the target more often when this object occurred in a high-variation scene than when it occurred in a low-variation scene. This was the case when the difference between the low- and high-variation conditions was large (like in Experiment 1) but also when this difference was more subtle (like in Experiment 2) or when the objects in the low-variation condition all had the same type (like in Experiment 3). In addition to this, a meta-analysis showed that these effects generalized over the different manipulations of scene variation in the low-variation conditions. Although the IA (Dale & Reiter, 1995) was designed as a computational interpretation of the Gricean Maxim of Quantity and uses a preference order that causes it to generate

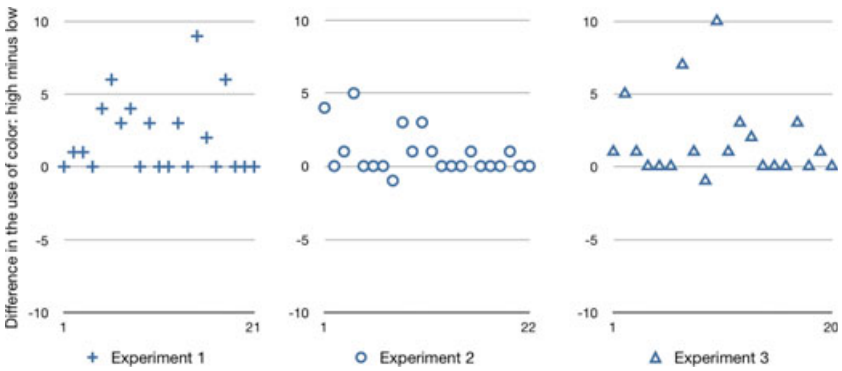

Fig. 8. Speaker variation in the three experiments: the difference in the use of color in the high- and the low-variation condition (*y*-axis) as a function of the individual speakers (*x*-axis). A positive value indicates that a speaker used color more frequently in the high- than in the low-variation condition.

target descriptions that might contain redundant attributes, it would never include color redundantly in any of the trials in the experiments discussed here. Our results show that speakers are sensitive to the amount of visual variation in a scene while the IA is not.

An interesting question, of course, is how an extension of the IA could account for the results presented here. We have seen that Dale and Reiter (1995) argue that type should always be included in a final description, even if does not rule out any distractors. Thinking along similar lines, one could consider extending the IA by allowing it to always include certain other attributes as well (such as color in high-variation scenes). In practice, this could be done by calculating a "variability index" for each particular scene, and deciding, based on this index, whether certain attributes should always be included or not. However, this solution seems rather ad hoc. For one thing, it would predict that speakers *always* include color in the high-variation condition, a prediction that clearly is not borne out by the data. Besides that, needless to say, visual variation is likely to be only one of a number of possible factors behind speakers' tendency to overspecify.

So what *do* the speakers in our experiments do? Even though our experiments were not set up to test specific models of the human production of referring expressions, we would like to discuss a search strategy that human speakers might use to decide about which attributes to (redundantly) include in their descriptions, related to the use of heuristics. Tversky and Kahneman (1974) argue that people rely on *quick heuristics* when making decisions, which they define as "beliefs concerning the likelihood of uncertain events (…) that reduce the complex tasks of assessing probabilities and predicting values to simpler judgmental operations" (p. 1124). It could well be that speakers also rely on heuristics when producing referring expressions, a suggestion that can also be found in Viethen and Dale (2009), Dale and Viethen (2009), and Van Deemter, Gatt, Van Gompel, and Krahmer (2012). Instead of carefully scanning a visual scene in search of target attributes with high distinguishing value, speakers might simplify the attribute selection process by using other strategies. Although our findings do not allow us to define specific heuristics that might be at play during reference production, we speculate that at least two interacting criteria play a role in this respect: visual saliency and scene gist.

Several studies have shown that speakers tend to include *visually salient* attributes (such as color) in their target descriptions, irrespective of their contrastive value (e.g., Belke & Meyer, 2002; Pechmann, 1989). This seems to be in line with speakers relying on heuristics when they need to describe a target in a scene with much variation: They will select visually salient attributes that immediately grab their attention, without making sure that these attributes do indeed help ruling out distractors. The same is the case with *scene gist*. On the basis of early work by Friedman (1979), Potter (1976), and in line with several other articles on the psychophysics of vision (e.g., Itti & Koch, 2001; Kremer & Baroni, 2010), Oliva (2005) defines the gist of a visual scene as the representation of that scene, including perceptual levels of processing (e.g., color, spatial characteristics) and conceptual levels of processing (e.g., objects, activation of semantic information). In line with this, Oliva, Torralba, Castelhano, and Henderson (2003) suggest that when people need to detect or describe a target object, they are initially guided by heuristics that result

from processing the scene on a perceptual level: People are inclined to take the visual context in which the target occurs into account. In this way, heuristics assist in focusing people's attention on relevant information about the target. Similarly, as describing objects requires detecting them first, the gist of a scene could also guide speakers' decisions on which target attributes to include in their referring expressions. For example, if a certain attribute in the context of the target attracts visual attention in some way, then it is more likely to be used as an attribute within a target description. Hence, if the variation in a scene is relatively high, it is likely that a speaker's attention is grasped by several attributes that vary across the objects in the scene. Given that in a high-variation scene, it will be less immediately obvious as to which attributes help in ruling out the distractors, speakers will be more likely to quickly include a property such as color. This will, at least in cases as those studied here, lead to overspecification.

One interesting question that remains is how scene variation affects listeners and what the role of redundant attributes would be in the target identification process. Regarding the latter, some articles claim that listeners are hindered by redundant target attributes during identification (e.g., Engelhardt et al., 2006), whereas others suggest the opposite and show that overspecification shortens the time needed for identification (e.g., Arts, Maes, Noordman, & Jansen, 2011). It seems plausible to assume that the amount of variation in the visual scene plays a crucial role here, as various studies have claimed there to be a close connection between scene perception and comprehension (see Ferreira and Tanenhaus [2007] for an overview). For one thing, in line with Engelhardt, Demiral, and Ferreira (2011), it could be the case that a redundant color attribute facilitates identification in a high-variation scene (because it could rule out one or more distractors), while it could possibly distract the listener in the case of a low-variation scene. Although this connection between scene variation and comprehension goes beyond the scope of the data presented in this article, we assume that it might cause problems for the current REG algorithms, as these are generally designed so as to mimic human speakers as much as possible. However, this does not necessarily cause them to generate descriptions that are of optimal use for the listener (Krahmer & Van Deemter, 2012), or that are adapted to the amount of variation in the visual scene in terms of their level of redundancy. Therefore, one challenge for the future would be to design algorithms that overcome these problems and to evaluate their output in terms of its benefits for the listener (see Gatt and Belz [2010] for a [rare] example of how this could be done).

7. Conclusion

This study demonstrates that speakers are more likely to include a color attribute in their target descriptions (even if this leads to overspecification) when the scene variation is high than when this variation is low. Our findings are problematic for existing algorithms that aim to automatically generate psychologically realistic target descriptions, such as the IA, as these algorithms make use of a fixed preference order per domain and do not take visual scene variation into account.

Acknowledgments

The research reported in this article forms part of the NWO VICI project, "Bridging the gap between psycholinguistics and computational linguistics: the case of referring expressions" (Grant 277-70-007). We thank Jette Viethen and two anonymous reviewers for their comments on an earlier version of this paper, and Tessa Dwyer and Joost Driessen for assistance in collecting the data. Parts of this paper have been presented at the 2011 Cognitive Science Conference in Boston, Massachusetts.

Note

1. These objects were taken from the Object Databank: http://www.tarrlab.org/.

References

Arts, A. (2004). *Overspecification in instructive texts*. PhD dissertation, Tilburg University. Nijmegen, the Netherlands: Wolf Publishers.

Arts, A., Maes, A., Noordman, L., & Jansen, C. (2011). Overspecification facilitates object identification. *Journal of Pragmatics*, *43*(1), 361–374.

Belke, E., & Meyer, A. (2002). Tracking the time course of multidimensional stimulus discrimination: Analysis of viewing patterns and processing times during "same" "different" decisions. *European Journal of Cognitive Psychology*, *14*, 237–266.

Dale, R. (1989). Cooking up referring expressions. In *Proceedings of the 27th annual meeting of the association for computational linguistics* (pp. 68–75). Vancouver, BC, Canada: University of British Columbia.

Dale, R. (1992). *Generating referring expressions: Building descriptions in a domain of objects and presses*. Cambridge, MA: MIT Press.

Dale, R., & Reiter, E. (1995). Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science*, *18*, 233–263.

Dale, R., & Viethen, J. (2009). Referring Expression Generation through attribute-based heuristics. In *Proceedings of the 12th European workshop on Natural Language Generation (ENLG)* (pp. 58–65). Athens, Greece.

Dale, R., & Viethen, J. (2010). Attribute-centric Referring Expression Generation. In E. Krahmer & M. Theune (Eds.), *Empirical methods in Natural Language Generation, lecture notes in computer science* (Vol. 5980, pp. 163–179). Berlin and Heidelberg: Springer.

Engelhardt, P. E., Bailey, K. G. D., & Ferreira, F. (2006). Do speakers and listeners observe the Gricean Maxim of Quantity? *Journal of Memory and Language*, *54*, 554–573.

Engelhardt, P. E., Demiral, S. B., & Ferreira, F. (2011). Over-specified referring expressions impair comprehension: An ERP study. *Brain and Cognition*, *77*, 204–314.

Ferreira, F., & Tanenhaus, M. K. (2007). Introduction to the special issue on language-vision interactions. *Journal of Memory and Language*, *57*, 455–459.

Friedman, A. (1979). Framing pictures: The role of knowledge in automized encoding and memory for gist. *Journal for Experimental Psychology: General*, *108*, 316–355.

Gatt, A., & Belz, A. (2010). Introducing shared task evaluation to NLG: The TUNA shared task evaluation challenges. In E. Krahmer & M. Theune (Eds.), *Empirical methods in Natural Language Generation* (pp. 264–293). Berlin and Heidelberg: Springer (LNCS 5790).

Gatt, A., Van der Sluis, I., & Van Deemter, K. (2007). Evaluating algorithms for the generation of referring expressions using a balanced corpus. In *Proceedings of the European workshop on Natural Language Generation (ENLG)* (pp. 49–56). Saarbruecken, Germany.

Gauthier, I., & Tarr, M. (1997). Becoming a Greeble expert: Exploring mechanisms for face recognition. *Vision Research*, *37*, 1673–1682.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Speech acts* (pp. 43–58). New York: Academic Press.

Griffin, Z., & Bock, K. (2000). What the eyes say about speaking. *Psychological Science*, *11*, 274–279.

Hanna, J., & Brennan, S. (2007). Speaker's eye gaze disambiguates referring expressions early during face-to-face conversation. *Journal of Memory and Language*, *57*(4), 596–615.

Itti, L., & Koch, C. (2001). Computational modeling of visual attention. *Nature Reviews Neuroscience*, *2*, 194–203.

Koolen, R., Gatt, A., Goudbeek, M., & Krahmer, E. (2011). Factors causing overspecification in definite descriptions. *Journal of Pragmatics*, *43*(13), 3231–3250.

Krahmer, E., & Van Deemter, K. (2012). Computational generation of referring expressions: A survey. *Computational Linguistics*, *38*(1), 173–218.

Kremer, G., & Baroni, M. (2010). Predicting cognitively salient modifiers of the constitutive parts of concepts. Proceedings of the Cognitive Modeling and Computational Linguistics workshop at ACL 2010, 54–62.

Mellish, C., Scott, D., Cahill, L., Evans, R., Paiva, D., & Reape, M. (2006). A reference architecture for Natural Language Generation systems. *Natural Language Engineering*, *12*(1), 1–34.

Meyer, A., Sleiderink, A., & Levelt, W. (1998). Viewing and naming objects: Eye-movements during noun-phrase production. *Cognition*, *66*, B25–B33.

Oliva, A. (2005). Gist of the scene. In L. Itti, G. Rees, & J. K. Tsotsos (Eds.), *Neurobiology of attention* (pp. 251–256). San Diego, CA: Elsevier.

Oliva, A., Torralba, A., Castelhano, M., & Henderson, J. (2003). Top-down control of visual attention in object detention. *Proceedings of the IEEE International Conference Image Processing*, *1*, 253–256.

Olson, D.R. (1970). Language and thought: Aspects of a cognitive theory on semantics. *Psychological Review*, *77*, 257–273.

Pechmann, T. (1989). Incremental speech production and referential overspecification. *Linguistics*, *27*, 89–110.

Potter, M. (1976). Short-term conceptual memory for pictures. *Journal of Experimental Psychology: Human Learning and Memory*, *2*, 509–522.

Reiter, E., & Dale, R. (2000). *Building Natural Language Generation systems*. Cambridge, UK: Cambridge University Press.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*, 1124–1131.

Van Deemter, K., Gatt, A., Van der Sluis, I., & Power, R. (2012). Generation of referring expressions: Assessing the Incremental Algorithm. *Cognitive Science*, *36*, 799–836. doi:10.1111/j.1551-6709.2011.01205.x

Van Deemter, K., Gatt, A., Van Gompel, R., & Krahmer, E. (2012). Toward a computational psycholinguistics of reference production. *Topics in Cognitive Science*, *4*(2), 166–183.

Viethen, J., & Dale, R. (2009). Referring Expression Generation: What can we learn from human data? In *Proceedings of the 2009 workshop on the production of referring expressions: Bridging the gap between computational and empirical approaches to reference*. Amsterdam, The Netherlands.