The impact of bottom-up and top-down saliency cues on reference production

Ruud Koolen (R.M.F.Koolen@uvt.nl)

Emiel Krahmer (E.J.Krahmer@uvt.nl)

Marc Swerts (M.G.J.Swerts@uvt.nl)
Tilburg Center for Cognition and Communication (TiCC), School of Humanities, Tilburg University
PO Box 90153, 5000 LE, Tilburg, The Netherlands

Abstract

This study investigates to what extent visual saliency cues in realistic visual scenes cause speakers to include a redundant color attribute in their definite descriptions of objects, and in particular how such cues guide speakers in determining which objects in the scene are relevant distractors, and which not. First, regarding bottom-up cues, the results revealed that the presence of clutter positively affected the redundant use of color, but that the distance between a target and a distractor did not have an effect in this respect. Second, an effect of top-down saliency (i.e., whether a target's type was mentioned in the instructions) was only partially borne out by the data. We argue that these findings are problematic for algorithms that aim to generate psychologically realistic object descriptions, since these generally select properties that help to distinguish a target from all distractors that are present in a scene.

Keywords: Definite reference; Overspecification; Bottom-up and top-down saliency; Computational models.

Introduction

When producing definite object descriptions (such as *"the green chair"*), speakers must decide on the information that they include in order to make a *target* object identifiable for the addressee. Many referential tasks require distinguishing a target from one or more *distractor* objects. The properties that speakers include to make the target identifiable seem to be largely determined by the properties of the distractor objects. For example, consider the two scenes depicted in Fig. 1.

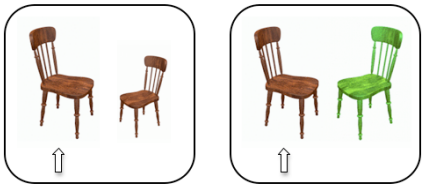

Figure 1: Two simple visual scenes.

Although the target is the same in both scenes (a large brown chair, as indicated by the arrows), the distractor will probably cause a speaker to describe it in different ways. In the left scene, where the distractor is a small brown chair, there is a high chance that a speaker produces a description like *"the large chair"*. However, in the right scene, where the distractor is a large green chair, a description like *"the brown chair"* is more likely to be uttered.

While the distractor object(s) seem to play a large role in the production of target descriptions in simple visual scenes such as the ones depicted in Fig. 1 (involving comparisons of structurally different minimal pairs of objects), it is the question whether a similar process is at play when speakers refer to target objects in realistic, more complex scenes. For example, imagine a speaker asking her listener to hand her a plate that is lying on a table full of objects. Do speakers then regard all these objects as relevant distractors? Or may there be reasons why certain objects are excluded from the set of distractors? And, most importantly, how does this influence the production of reference? These are the questions that we address in the current paper.

Background

In recent years, the production of referring expressions has received considerable attention, both from a computational and from a psycholinguistic perspective (van Deemter et al., 2012). In computational linguistics, for instance, researchers have developed several Referring Expression Generation (REG) algorithms, most notably the Incremental Algorithm (IA) introduced by Dale and Reiter (1995). The IA is a computational model that focuses on *content planning*: the algorithm iteratively selects attributes (e.g., type, color, size) in order to distinguish a target from one or more distractor objects in the *distractor set*. In order to do this, the IA uses a preference order that contains all attributes that occur in the given domain, where it considers frequently used attributes for inclusion before less frequent attributes. In this paper, we assume that *type* is at the head of the preference order (before *color*), since it is needed to generate a proper noun phrase (Levelt, 1989).

So how does the IA define the distractor set? Dale and Reiter (1995) write: "We define the context set to be the set of entities that the hearer is currently assumed to be attending to" (p. 236), where the distractor set consists of all elements that are present in a visual scene except the target. Thus, Dale and Reiter do not explain explicitly how the set of distractors should be determined for a scene, and whether it should be restricted in a certain communicative situation or not. This means that the IA generally selects the content of its object descriptions by searching for properties that help to distinguish the target from *all* distractors that are present in a visual scene. This might be problematic from a psychological perspective: for example, regarding *discourse*

structure, Krahmer and Theune (2002) argue that the set of distractor objects may change during a discourse (e.g., when speakers repeatedly refer to the same object), while Kelleher and Kruijff (2006) and Van der Sluis (2005) add that *visual salience* may play a role in this as well. However, to the best of our knowledge, earlier work that systematically tests how human speakers are driven by visual salience to dynamically restrict the distractor set for a given scene is lacking.

In the current paper, we test three visual saliency cues that may guide speakers in determining the set of distractors. We base our manipulations on Itti and Koch's (2000) model of visual attention, stating that an object pops out of a scene if it is sufficiently salient (Itti and Koch call this *bottom-up*, perceptual saliency), or if the viewer's attention is guided to it (also referred to as *top-down*, conceptual saliency).

With regard to bottom-up scene processing, we expect two kinds of cues to guide speakers in determining the distractor set. First, we expect the presence of *visual clutter* to play a role. We define visual clutter here as a collection of objects that are thematically related to the target object, and assume that the amount of visual clutter is positively correlated to the amount of objects in a scene (Bravo & Farid, 2008). In previous research, clutter has been shown to affect speakers' response times when describing naturalistic scenes (Coco & Keller, 2009), with slower reactions for cluttered scenes. In line with this, we expect that since a cluttered scene contains more objects (and may thus be more difficult to process), it is unlikely that speakers 'calculate' for every distractor how it can be distinguished from the target object.

Secondly, again regarding bottom-up scene processing, we expect *distractor distance* (that is, the distance between the target and a distractor) to guide speakers in determining the distractor set. For reference in dialogue, Beun and Cremers (1998) suggest that a speaker's *focus of attention* limits the number of relevant distractors: in their experiment, they find that speakers generally consider only visually close objects when referring to targets. In an eye-tracking study, Brown-Schmidt and Tanenhaus (2008) have similarly shown that distractors that are visually close to the last mentioned target are most likely to be in the speaker's focus of attention. In this paper, we study if the same goes for reference where no preceding discourse is involved.

Thirdly, related to top-down scene processing, we expect the *specificity of the referential task* to affect speakers when determining the set of distractors, where we hypothesize that a general task (such as "describe this object") will leave the speaker with a bigger, less restricted set of distractors than a more specific task (such as "describe this plate", where the target's type is mentioned). In the latter case, speakers might leave objects other than plates unattended, while any object that is present in the scene might be regarded as a relevant distractor in the former case.

The current study

Our experiment was a reference production task, in which participants were presented with realistic scenes on a screen. The scenes contained one target and several distractors, and the participants were asked to describe the target in such a way that an addressee could uniquely identify it. Crucially, the trials were set up in such a way that color was never needed to do this, enabling us to take the proportional use of *redundant* color attributes (i.e., color attributes that were not necessary for identification) as our dependent variable. In doing this, we follow Koolen et al. (to appear), who used the redundant use of color to study how speakers differ in their perception of low-variation and high-variation scenes. Our stimuli were designed in such a way that the Incremental Algorithm would *not* select color to distinguish the target: it would always select *type* and *size*, irrespective of any experimental condition.

What do we predict regarding the redundant use of color? Firstly, we expect to find an effect of *clutter* (with speakers using more color when clutter objects are present), because a cluttered scene contains more distractors and hence might be more difficult to process. Secondly, we expect *distractor distance* to affect speakers' redundant color use: redundant color attributes might get used more often when a potential distractor is placed close to the target object as compared to when it is distant. Thirdly, we expect to find effects of the *specificity of the referential task*: we hypothesize that when speakers are instructed in a general way (e.g., "Describe this object"), they will be more likely to redundantly use color as compared to when the instruction includes the target's type (e.g., "Describe this plate").

Experiment

Method

Participants. 43 undergraduate students from Tilburg University (30 female, 13 male) took part in the experiment. All (mean age 21 years and 1 month, range 18 - 34 years) were native speakers of Dutch (the language of the study) and participated for course credits.

Materials. The stimulus materials consisted of 80 trials, all of which were photo-realistic pictures of objects on either a kitchen table or an office desk. In the 40 critical trials, there were always at least three objects present on the table: one target object and two distractor objects. Crucially, one of the distractors had the same type and color as the target object (meaning that it could only be ruled out by means of its size), and was always positioned next to the target object (either left or right). The second distractor always had a different type and color as compared to the target, and its positioning varied across conditions. Besides that, two other principal factors were manipulated: one related to the presence of clutter in the scene, and one related to the specificity of the task that was given to the speakers.

The first manipulation was related to whether or not there was *clutter* present in the visual scene. We define clutter as a collection of all kinds of objects that are thematically related to the target and its two main distractors. The clutter objects were not systematically varied, and were unique in the sense that they all had different types. The color of these

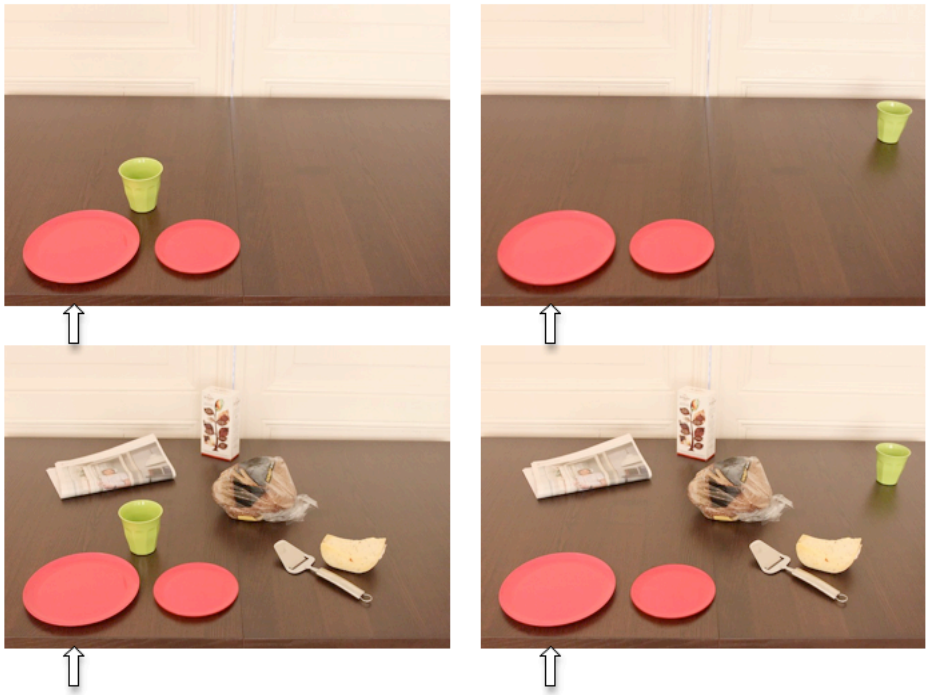

Fig. 2: Examples of critical trials in Experiment 1. The left scenes have a close distractor, whereas this distractor is distant in the right scenes. The upper scenes do not contain clutter, whereas the lower scenes do. Note that both the small and the large plate could be the target in the experiment.

clutter objects was kept as neutral as possible; it was at least made sure that the clutter objects did not have the same color as compared to the target and its two distractors. In the cluttered pictures in Fig. 2, five objects are added that one would expect to see on a breakfast table, where most do not have a salient color: a bag of bread, a newspaper, a piece of cheese, a cheese slicer, and a pack of chocolate sprinkles. Clutter was added in half of the critical trials, and the same clutter objects were used for the scenes with a close and a distant distractor.

The second manipulation (*distractor distance*) was related to the distance between the target object and the second distractor. This distance was manipulated as follows: in half of the trials, the distractor was positioned close to the target (with the two distractors placed in the same corner of the table), whereas this distance was maximized in the other half of the trials (with the target and the first distractor in one corner of the table, and the second one in the opposite corner). In Fig. 2, the left pictures had a *close* distractor, and the right pictures had a *distant* distractor. When a scene had a distant distractor, this object was always positioned in the corner opposite the target. Note that mentioning the target's type and size was sufficient to identify the target in both the close and distant conditions, implying that the use of color would inevitably result in overspecification.

The experiment had eighty trials: forty critical trials and forty fillers. Regarding the critical trials, we used ten scenes: five scenes with objects on an office desk and five scenes with objects on a kitchen table. These ten scenes were all manipulated in a 2 (*distance*) x 2 (*clutter*) design, resulting in four within-conditions as described above: one picture with two close distractors but without clutter, one with a close and a distant second distractor without clutter, one with two close distractors and with clutter, and one with a close and a distant distractor and with clutter. Note that the target object could be positioned in all four corners of the table (and not necessarily in the left bottom corner, as is the case in Fig. 2). Since there were always two similar objects in a scene (one of which being the target object), we marked the small object as the target in half of the scenes, and the large object in the other half of the scenes.

Besides distractor distance and the presence of clutter presence (both manipulated as within participants factors), the experiment also had one between participants variable (hence called *specificity of the referential task*), which was related to the instruction that was given to the participants. As mentioned earlier, it was the participants' task to describe each target object in such a way that it could be distinguished from the other objects in the visual scene. All participants were presented with the same stimuli, but two kinds of instructions were used. Half of the participants had the task to "describe this object" (which means that they took part in the *low specificity condition*), whereas the other half of the participants (in the *high specificity condition*) had a more specific task. In this condition, the target's type was mentioned in the instruction. For example, in Fig. 2, these speakers were asked to "describe this plate".

Table 1: Overview of the experimental design and the number of descriptions within each cell.

	No clutter		Clutter	
	Close	Distant	Close	Distant
Low specificity	210	210	210	210
High specificity	220	220	220	220

The experiment had forty fillers: twenty from the kitchen table domain and twenty from the office desk domain. These fillers were set up in the same way as the critical trials, in the sense that there were scenes containing few objects that were positioned in the same way as those in the critical trials, and scenes containing many different objects (in line with the clutter scenes that served as critical trials). Again, one of the objects was marked as the target and was described by the participants, with the crucial difference that the objects in the filler pictures did not differ in terms of their color. In this way, speakers were discouraged from using color when describing the fillers.

Procedure. The experiment was performed in a lab, and had an average running time of 10 minutes. After participants had entered the lab, they were randomly assigned to one of the two conditions: 21 participants took part in the low specificity condition, and 22 in the high specificity condition. Thereafter, they were seated opposite the listener (who was a confederate of the experimenter), and were instructed so as to describe a target in such a way that their listener could uniquely identify it. For each new trial, the participants were instructed by means of a pre-recorded task (for example: "describe this object" in the low specificity condition and "describe this plate" in the high specificity condition). Speakers could take as much time as needed to describe a target, and their descriptions were recorded with a voice recorder.

The trials were presented to participants on a computer screen. We made one block of eighty trials in a fixed random order (which was presented to one half of the participants), and a second block containing the same trials in the reverse order (which was presented to the other half of the speakers). There were two practice trials. The listener had a paper booklet in front of her, containing - for each trial - separate pictures of all the objects that occurred in that given scene. These pictures were taken from the pictures the speaker was presented with. Based on the speaker's descriptions, the listener marked the object that she thought was referred to on an answering form. In order to prevent speakers from including location information in their target descriptions (e.g., *'The plate in the left bottom corner'*), the instructions emphasized that the listener was presented with the same objects ranked in a different order. The listener always acted as though she understood the descriptions, and never asked clarification questions. This was done to enable a focus on content planning of initial descriptions ('first mentions'). Once the listener had identified a target, this was communicated to the speaker, who then went on to describe the next target.

Design and statistical analysis. The experiment had a 2 x 2 x 2 design (see table 1) with two within participants factors: *distractor distance* (levels: close, distant) and *clutter presence* (levels: no clutter, clutter), and one between participants factor: *specificity of the referential task* (levels: low, high). The experiment had one dependent variable: the proportion of descriptions containing a color attribute. As described above, we made sure that speakers never needed color in order to distinguish the target object from its distractors: mentioning the target's type and size was always sufficient. Thus, when speakers mentioned color, this always resulted in an overspecified description.

Our statistical procedure consisted of Repeated Measures ANOVAs: one on the participant means (F_1) and one on the item means (F_2). We only report on interactions where these are significant. In order to compensate for departures from normality, we applied a standard arcsin transformation to the proportions before running the ANOVAs. For the sake of readability, we report the untransformed proportions in the results section.

Results

In total, 1720 target descriptions were produced in this experiment. All of these contained a type attribute, and most (85.8%) contained a size attribute. In the rest of the cases, other additional attributes were mentioned to distinguish the target object (such as its orientation). All descriptions were fully distinguishing. Speakers redundantly mentioned color in 39% of the descriptions.

Results for clutter. The first factor that we expected to affect speakers' redundant use of color was the presence of visual clutter in the scene. Fig. 3 displays the proportional use of color as a function of clutter presence.

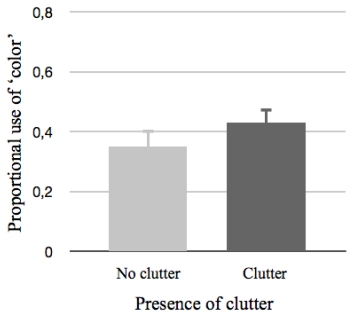

Fig. 3: The proportional use of color (plus standard deviations) as a function of clutter presence.

As can be seen in Fig. 4, the presence of clutter positively affected the redundant use of the attribute color ($F1_{(1,41)}$ = 13.38, p = .001; $F2_{(1,36)}$ = 3.91, p = .06). In other words, speakers were more likely to include color when presented with visual scenes containing clutter (M = .43, SD = .05) as compared to when the scene did not contain clutter (M = .35, SD = .06).

Results for distractor distance. We also studied the effect of distractor distance on the redundant use of color: whether a distractor was placed close to or far from the target object. The results are shown in Fig. 4.

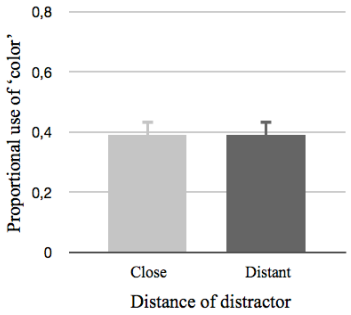

Fig. 4: The proportional use of color (plus standard deviations) as a function of distractor distance.

As can be seen in Fig. 3, distractor distance did not affect the proportional use of the redundant attribute color ($F1_{(1,41)}$ = .068, p = .80; $F2_{(1,36)}$ = .00, p = .99). More specifically, color was mentioned color exactly as many times when the distractor was close (M = .39, SD = .05) as compared to when it was distant (M = .39, SD = .05).

Results for specificity of the referential task. The third factor that we manipulated was related to the specificity level of the task that was given to the speakers: for half of the speakers this specificity level was low, while it was high for the other half of the speakers.

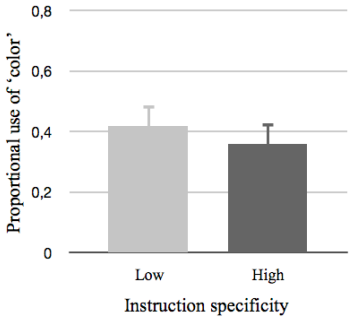

Fig. 5: The proportional use of color (plus standard deviations) as a function of the specificity of the instruction.

As reflected in Fig. 5, the specificity of the instructions to some extent affected the use of the redundant attribute color, but this effect was only significant by items ($F1_{(1,41)}$ = .355, p = .55; $F2_{(1,36)}$ = 15.81, p < .001). More specifically, this means that speakers that took part in the low specificity condition (M = .42, SD = .07) did use color more frequently as compared to those taking part in the high specificity condition (M = .36, SD = .07), but that we did not find a convincing effect of instruction specificity.

Discussion

In this paper, we have investigated how bottom-up and top-down saliency cues - as defined by Itti and Koch (2000) - guide speakers in determining which objects in a scene belong to the set of relevant distractor objects. In doing this, we have studied how these cues affect speakers' production of object descriptions, and in particular, to what extent they cause speakers to mention a redundant color attribute. On average, 39% of the object descriptions in our experiment redundantly contained a color attribute, which is more than the proportions reported by, among others, Belke and Meyer (2002), Koolen et al. (to appear), and Pechmann (1989).

Regarding *top-down* scene processing, we hypothesized that participants in the low specificity condition (who were asked to "describe this object") would be more likely to use a redundant color attribute than participants that took part in high specificity condition (who were asked to "describe this X", e.g., "this plate"), since in the latter case (where the target's type was used in the instruction) only the distractor with the same type would remain to rule out (which could always be done by mentioning size). We indeed found a numerical difference between the conditions in the predicted direction, but this was only significant in the F_2 analysis. We plan to further study this effect in future research.

Secondly, regarding bottom-up scene processing, we have found that - at least for the visual scenes used here - the *distance* between the target and a distractor does not affect the redundant use of color. This might be due to an artefact of the experimental setup: given that our speakers knew that the addressee was presented with – for every given scene – separate pictures of all objects that were depicted in that scene, this might have caused them to ignore the distance between the target and the distractors in the scene. In future research, we aim to improve our manipulation of distractor distance.

Thirdly, again regarding *bottom-up* scene processing, our results showed that speakers are more likely to mention a redundant color attribute when there is *visual clutter* present in the scene as compared to when this is not the case. One explanation for this might be that a scene with clutter simply contains more distractor objects than a scene without clutter. As suggested earlier, the latter might lower the chance that speakers exactly 'calculate' for each distractor how it can be distinguished from the target in the most efficient way. Our results suggest that speakers tend to process cluttered scenes in a 'faster' way: they might rely on *heuristics* (Tversky & Kahnemann, 1982) when they have to uniquely describe an object. With regard to reference production, heuristics can be defined as general rules that say, for example, that color

should always be included in the case of a cluttered scene. In recent years, such heuristics have indeed been claimed to influence reference production (e.g., Dale & Viethen, 2009), since speakers' limited processing capacity might prevent them from calculating the shortest possible description in a given referential task (van Deemter et al., 2012).

As we have explained in the introduction of this paper, for the current REG algorithms (most notably Dale and Reiter's IA introduced in 1995) it is not explained explicitly how the distractor set should be defined for a given scene (Krahmer & Theune, 2002): such algorithms select the content of their descriptions by searching for those properties that help to distinguish the target from *all* distractors that are present in the scene. Given that our findings (at least partly) suggest that perceptual and conceptual cues affect speakers' object descriptions, and that there are situations in which speakers do not take all distractors into account, or do not 'calculate' the most efficient way to describe a target because too many objects are present, the question is what the implications of our findings are for REG algorithms such as the IA.

For one thing, our results show that speakers often use a color attribute when algorithms such as the IA would not do this: in our stimuli, the algorithm would select *type* and *size* instead of *color* (assuming that, as explained earlier, *type* is placed at the head of the preference order, followed by *color* and other, less preferred attributes such as size). So how can the IA account for this frequent color use? For the specific case of clutter (which delivered us with the most convincing results), we propose that one solution might be to make the algorithm redundantly include color more often when it has to describe a target object in a cluttered scene as compared to a scene without clutter. For example, this could be done by dynamically adapting the preference order to the amount of clutter that is present in a particular scene: when clutter is present, color could be placed at the head of the preference order (causing it to be selected if there is any color variation in the scene), whereas the preference order can remain as we assume it is now (with *type* before *color*) for visual scenes that do not contain clutter.

Conclusion

Bottom-up cues (i.e., the presence of visual clutter, but not distractor distance), and to a lesser extent top-down saliency cues (i.e., specificity of the referential task) influence the redundant use of color in definite object descriptions. This is problematic for Referring Expression Generation algorithms that aim to automatically generate psychologically realistic object descriptions.

Acknowledgments

The research reported in this paper is 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 Anouk van Helvoort for help in creating the stimuli, Kristel Bartels and Elsa Jonkers for running the experiment and annotating the data, and Jette Viethen for comments on an earlier draft of this paper.

References

Beun, R., & Cremers, A. (1998). Object reference in a shared domain of conversations. *Pragmatics and Cognition*, 6 (1/2), 121-152.

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.

Bravo, M. & Farid, H. (2008). A scale invariant measure of clutter. *Journal of Vision*, 8 (1), 1-9.

Brown-Schmidt, S., & Tanenhaus, M. (2008). Real-time investigation of referential domains in unscripted conversation: a targeted language game approach. *Cognitive Science*, 32 (4), 643-684.

Coco, M., & Keller, F. (2009). The impact of visual information on reference assignment in sentence production. In *Proceedings of the 31st annual conference of the Cognitive Science society*, 274-279. Amsterdam, The Netherlands.

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

Dale, R., and Viethen, J. (2009). Referring Expression Generation through attribute-based heuristics. *Proceedings of the 12th European workshop on NLG (ENLG)*, 58-65.

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision research*, 40, 1489-1506.

Kelleher, J. & Kruijff, G.J. (2006). Incremental generation of spatial referring expressions in situated dialogue. In *Proceedings of COLING/ACL '06*. Sydney, Australia.

Koolen, R., Goudbeek, M., and Krahmer, E. (to appear). The effect of scene variation on the redundant use of color in definite object descriptions. *Cognitive Science*.

Krahmer, E., & Theune, M. (2002). Efficient context-sensitive generation of referring expressions. In: K. van Deemter & R. Kibble (Eds.). *Information sharing: Givenness and newness in language processing* (pp. 223-264). CSLI publications, Stanford.

Levelt, W. (1989). Speaking: from intention to articulation. MIT Press, Cambridge/London.

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

Tversky, A. & Kahneman, D. (1982). Judgement under uncertainty: heuristics and biases. In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgement under uncertainty, heuristics and biases*. Cambridge: Cambridge University Press.

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.

Van der Sluis, I. (2005). Multimodal reference. *PhD thesis*, Tilburg University, The Netherlands.