Effects of domain size during reference production in photo-realistic scenes

Koolen, Ruud; Krahmer, Emiel

Published in:
Proceedings of the 42nd Annual Conference of the Cognitive Science Society

Publication date:
2020

Document Version
Peer reviewed version

Effects of domain size during reference production in photo-realistic scenes

Ruud Koolen (r.m.f.koolen@tilburguniversity.edu)
Tilburg Center for Cognition and Communication (TiCC); Department of Communication and Cognition
Tilburg University, The Netherlands

Emiel Krahmer (e.j.krahmer@tilburguniversity.edu)
Tilburg Center for Cognition and Communication (TiCC); Department of Communication and Cognition
Tilburg University, The Netherlands

Abstract

The current study investigates how speakers are affected by the size of the visual domain during reference production. Previous research found that speech onset times increase along with the number of distractors that are visible, at least when speakers refer to non-salient target objects in simplified visual domains. This suggests that in the case of more distractors, speakers need more time to perform an object-by-object scan of all distractors that are visible. We present the results of a reference production experiment, to study if this pattern for speech onset times holds for photo-realistic scenes, and to test if the suggested viewing strategy is reflected directly in speakers' eye movements. Our results show that this is indeed the case: we find (1) that speech onset times increase linearly as more distractors are present; (2) that speakers fixate the target relatively less often in larger domains; and (3) that larger domains elicit more fixation switches back and forth between the target and its distractors.

Keywords: Reference; language production; domain size; eye movements; speech onset times.

Introduction

Suppose you want to point out the marked object in Figure 1 to a listener. To fulfil this task, you would probably produce a definite referring expression such as *"the large blue plane"*, to distinguish the target referent from its surrounding distractors; in this case the other planes in the scene. Determining the content of such descriptions, which are ubiquitous in everyday language, requires visual inspection of the scene and the objects therein. What viewing strategy would you use?

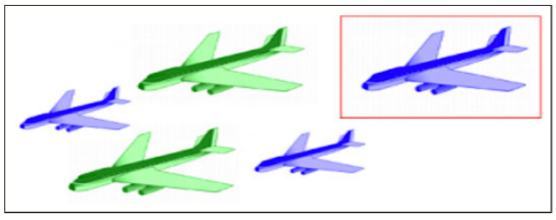

Figure 1: An abstract visual scene, taken from Gatt, Krahmer, Van Deemter, and Van Gompel (2017).

You could decide to carefully scan *all* planes in the scene separately, to make sure that your description is fully distinguishing. With such a strategy, you would follow the majority of current algorithms for automatic Referring Expression Generation (REG; see Krahmer and Van Deemter, 2012, for a review), which model content selection as a search for (a combination of) attribute that are unique for the target object at hand. Arguably, applying such a systematic visual search could work for human speakers as well, in particular when referring to targets in simpler visual scenes with a relatively small number of distractors. In such a scenario, one would expect that speakers' search times would increase along with the number of distractors (hence referred to as *domain size*), echoing the intuitive assumption that comparing more objects simply takes more time. As explained below, this assumption has been confirmed by Gatt, Krahmer, Van Deemter, and Van Gompel (2017), at least under certain conditions.

In their experiments, Gatt et al. (2017) manipulated visual scenes like the one depicted in Figure 1: they had participants refer to targets in rather artificial scenes consisting of separate objects in a grid. This approach raises the question how their findings extend to photo-realistic scenes: would search times depend on domain size in a similar way there, or would patterns change due to the increased realism? And how are these patterns reflected in speakers' inspection of the visual scene? These questions are central to the current paper.

Theoretical background and relevance

Over the years, a substantial body of research has shown that vision and language are closely intertwined. From a language comprehension perspective, the dominant approach has been to uncover the link between visual and linguistic processing using the Visual World Paradigm (Tanenhaus, Spivey, Eberhard & Sedivy, 1995). For language production, the challenge has been – and still is – to explore how a large variety of visual scene characteristics shape the planning and realization of, for example, definite referring expressions. In some cases, previous work in this direction based its predictions on visual perception studies. For example, the robust finding that color contrast - which is an instance of low-level visual information - causes speakers to include color in their object descriptions (Koolen, Goudbeek & Krahmer, 2013; Rubio-Fernández, 2016) is well-reflected in (even early) models of visual attention: already in 1980, Treisman and Gelade showed that color differences "pop-out" of the scene.

While this example with color is rather intuitive, the vision literature can also be a good place to start when making predictions about how variables such as domain size could affect reference production. In visual search, the traditional task for participants is to confirm or deny the presence of a target in a visual display (Wolfe, 2010). Within this body of research, domain size is referred to as set size, and it is measured how

the number of distractors affects search times. Also here, we find the pop-out effect to play a role: when searching for targets that are somehow salient in a display, set size does hardly affect search times, while a linear increase in search times is observed as a function of set size for targets that fail to pop out (e.g., Palmer, 1995; Treisman & Gelade, 1980). The latter has for example been shown to occur when the number of properties shared by the target and the distractors is relatively high (Nordfang & Wolfe, 2014).

Interestingly, in two recent experiments, Gatt et al. (2017) show that these findings from the vision literature can be extended to reference production as well. Gatt et al. had participants refer to target objects in visual scenes with either 2, 4, 8 or 16 distractors, and measured how speech onset times varied as a function of domain size. The results seem to resemble the ones from visual search studies: in scenes where the target popped out, and could be distinguished by means of its color, speech onset times were not affected by domain size. However, when the target did not pop out, and size was one of the distinguishing properties, onset times increased more or less linearly as a function of the higher number of objects.

With their experiments, Gatt and colleagues (2017) provide empirical evidence for the occurrence of the pop-out effect in reference production, echoing the robust finding that absolute (color) attributes are more often included than relative (size) attributes (e.g., Belke, 2006; Pechmann, 1989). The explanation that they give for their results is that absolute properties do not require an extensive one-by-one scan of the distractors in the scene, while relative properties do, and that an exhaustive scan simply takes longer in larger domains. Although this explanation is plausible, Gatt et al.'s (2017) experiments raise interesting follow-up questions. Two of them are addressed in the current paper: one on the use of photo-realistic scenes; and one on the use of eye tracking data to take a direct measure of scene perception as a function of domain size.

Rather than photo-realistic visual scenes, Gatt et al. (2017) manipulated abstract scenes that consisted of separate objects placed in a grid, without giving any further context (see again Figure 1). Such simplified visual scenes are commonly used in experiments on human and automatic reference production (e.g., Koolen et al., 2013; Rubio-Fernández, 2016; among many others), since they allow for controlled stimulus materials and manipulations. However, when looking at the vision literature, we see that in more complex realistic scenes, viewers do not solely rely on low-level visual features, but also on more semantic factors, such the prior expectations and understanding they have of the scene (e.g., Henderson, Brockmole, Castelhano & Mack, 2007; Torralba, Oliva, Castelhano & Henderson, 2006).

To test if the effect of set size changes when realistic scenes are used, Neider and Zelinsky (2008) report on a visual search experiment that manipulated set size in near-realistic scenes. The authors found the reverse effect of earlier work that used more artificial grids (e.g., Palmer, 1995; Treisman & Gelade, 1980; Nordfang & Wolfe, 2014): search got more efficient as more objects were present. Neider and Zelinsky explain this result by arguing that a higher number of objects in realistic scenes facilitates search, since it helps viewers to understand what the scene is about, and to restrict the set of distractors to a functional set. In other words: viewers know better what to fixate (and what not), and are not so much guided by salience-based pop-out mechanisms (Henderson, Malcolm & Schandl, 2009). For example, when searching for a pen in a scene of a living room, viewers will scan the objects on the table surface rather than, say, the objects on the couch. This way, increased realism interferes with the effect of set size: it restricts object-by-object searches to smaller parts of the scene.

The question is whether using photo-realistic scenes (rather than abstract grids) has a similar impact during reference production: would it change the effect of domain size on speech onset times, and if so, how? One prediction could be that the realistic setting reduces the impact of salience-based pop-out mechanisms, thus requiring an object-by-object scan of the distractors in the scene. In that scenario, one predicts to find a linear increase in onset times as the domain size get bigger, akin to Gatt et al.'s (2017) findings for targets that fail to pop out. On the other hand, if photo-realistic scenes allow speakers to rely on a set of only functional distractors, speech onset times may not be affected by domain size so much, or even decrease in bigger domains. The current experiment aims to provide empirical evidence for either one of these predictions, by manipulating photo-realistic photographs depicting groups of people. The photos are presented to speakers in both color and black-and-white, since the absence of color may even further reduce the chance of the pop-out effect to occur (e.g., Henderson & Hollingworth, 1998).

Since speech onset times cannot tell us directly what search strategy speakers apply as a function of domain size, we also measure speakers' eye movements. The use of eye tracking to study visually-grounded reference production has become increasingly popular in recent years (e.g., Davies & Kreysa, 2017; Elsner, Clarke & Rohde, 2018). Regarding domain size, one prediction could be that not only speech onset times increase as there are more objects in the scene, but also that a big domain size leads to: 1. fewer fixations on the target rather than the distractors; 2. more fixation switches between the target its distractors. On the other hand, if realistic scenes allow speakers to rely more on a set of functional distractors, speakers' fixations and selected attributes may not depend on domain size so much, or may show the opposite patterns like the ones sketched above.

Method

We performed a reference production experiment that asked participants to produce spoken target descriptions of people depicted in realistic photographs. We recorded speech as well as eye movements, to make a link between visual processing and speech onset times.

Participants

Participants were 58 undergraduate students from Tilburg university, who earned course credits for taking part in the experiment. There were two criteria for participants to take part: they should be native speakers of Dutch, which was the

language of the study; and they should not be wearing glasses or contact lenses, since these may give problems when calibrating the eyes with the eye tracker. Still, data for 14 participants could not be analyzed, primarily because their eye movements failed to record (note that we could not use a head rest since participants also had to speak), or did not evolve in a natural way. For example, if there was hardly any variation between gaze positions for a trial, we knew that something had gone wrong with the calibration. The data for these participants was discarded, on top of the data for some participants who produced extensive rather than distinguishing target descriptions. The actual data set consisted of data for 44 participants (24 female, 10 male; age range 18 - 27 years old; M = 22 years and 2 months) that was useful for further analysis.

Materials

A basic set of eight realistic photographs in full color served as the starting point for the creation of the critical trials. These photographs were collected on the web: we did a Google search for photographs (hence called: scenarios) depicting a group of at least seventeen people, standing or sitting in front of rather neutral backgrounds. The neutral backgrounds allowed us to manipulate domain size within the same scenario: by using a program called Impaint, we could 'erase' people from the scenes, creating trials with three, five, and nine people, based on the original scenarios with seventeen people. These numbers of objects represented the manipulation of our first independent variable, *domain size*. The levels of this variable corresponded to the levels of domain size applied by Gatt et al. (2017), who also had trials with one target referent and either two, four, eight, or sixteen distractor objects.

Our manipulations of the eight basic scenarios resulted in 32 photographs in full-color that could be used for the experiment: eight scenarios for all four levels of domain size. Critical trials were then created by adding red arrows pointing at one of the three persons that were present in all trials, irrespective of domain size (i.e., the three people that were left in the eight trials with the smallest domain size). This way, 96 (32 x 3) unique critical trials were created, all with one marked target person that had to be referred to by our participants. The people surrounding this target served as distractor objects in the reference production task. While the original scenarios were all in full color, we also created black-and-white versions for all of them, to manipulate the second independent variable of the research: *saturation*. Although these trials were in black-and-white, the arrows marking the target persons within these trials scenes were still in red, to make them sufficiently salient. Figure 2 depicts the eight critical trials that were created on the basis of one basic scenario, representing all four domain sizes in color versions.

Every participant saw a subset of the total number of critical trial pictures that was available. Four subsets of were created, representing four lists of critical trials (both in color and black and white). Essentially, every list consisted of six repeated items for all four levels of domain size, which makes 24 critical trials for the whole experiment. We made sure that every list contained every basic scenario three times, but always with different domain sizes, and different target persons. For example, for the basic scenario in Figure 2, the first list con-

Figure 2: Examples of critical trials in the color condition, representing the four levels of Domain size: two distractors (upper left), four distractors (upper right), eight distractors (lower left), and sixteen distractors (lower right).

tained the trials with two, four, and eight distractors (with three times a different target person); the second list contained the trials with sixteen, two, and four distractors (again with three different target persons); and so forth. This way, we made sure that participants never referred to the same unique target person more than once, to avoid any possible interference with our manipulation of domain size.

In all lists, the 24 critical trials were mixed with 24 filler items. The fillers consisted of three Greebles (Gauthier & Tarr, 1997) that were depicted next to each other. Greebles are abstract 3D figures, which could be distinguished from each other by means of their main shape or by the direction in which their protrusions were pointing. As such, they elicited the use of attributes that were different from the attributes that were generally used to describe the target referents (i.e., humans) in the critical trials, which distracted our participants from our manipulations. The combination of critical trials and Greeble fillers resulted in a total of 48 trials for every participant. The order in which these trials were presented was not random, since we wanted to be sure that participants did not see, say, three trials in a row representing one particular domain size or basic scenario; or, say, five Greeble fillers in a row. Therefore, there were two fixed orders for every list: one pre-determined order, and the corresponding reverse order.

Procedure

The experiment collected spoken Dutch target descriptions, and took place in our laboratory at Tilburg university. Before data collection started, we pre-registered our study at the Open Science Framework (OSF; www.osf.io). All participants signed a written informed consent form, which was approved by the ethics committee of the Tilburg Center for Cognition and Communication (Tilburg University). The consent form contained a general description of the experimental task, an indication of the duration of the experiment, contact information, and information about data storage. All participants gave explicit permission to have their audio recordings and eye movement data used for research purposes. During the task, participants were allowed to quit the experiment at any stage; none of the participants decided to do so. It took around 20 minutes to complete the experiment.

After entering the lab, participants were seated in a soundproof booth, and read and signed the consent form. They were then presented with an elaborate instruction, on paper, which explained that it was the participants' task to produce oral descriptions of target referents in visual scenes (being both people and abstract objects), in such a way that these targets could be distinguished from the surrounding distractor objects or people in the scene. Furthermore, participants were instructed to avoid location information in their descriptions (e.g., *the girl in the left bottom corner*). After reading the instruction, there was room for questions. The next step was to calibrate the eyes of the participant to the eye tracker, for which a 9-point validation method was used. Once the calibration was successful, participants completed three practice trials, for which the data was not analyzed. After this practice stage, there was a another (final) opportunity to ask questions about the task. The experimenter left the soundproof booth right before the start of the actual experiment.

All participants completed a total of 48 trials (24 critical trials and 24 fillers). They were randomly assigned to either the color or the black-and-white condition, and to one of the four corresponding lists (note that Saturation was manipulated between participants). The 44 participants whose data was included in the final analysis were equally divided over the two Saturation conditions. Eye movements were measured with an SMI RED 250 device, operated by the IviewX and the ExperimentCenter software packages. The eye tracker had a sampling rate of 250 Hz. The viewing distance was 70 cm. A headset microphone was used to record the participants' descriptions.

The stimuli were displayed on a 22 inch P2210 Dell monitor, with the resolution set to 1680 x 1050 pixels. The images were depicted in the middle of the screen, and resized to 1267 x 950 pixels (without changing the aspect ratio), surrounded by grey borders. These borders were required since eye tracking measurements outside the calibration area (which almost covered the whole screen, but not its most peripheral areas) are not fully reliable. Before trial onset, a fixation cross appeared in the centre of the screen. By fixating this cross for one second, the next trial appeared automatically. When fixating the cross did not work, participants could continue to the next trial manually by pressing spacebar.

Research design and data annotation

The experiment had a 4 x 2 mixed design, with Domain size (levels: 2, 4, 8, 16 distractors) manipulated within participants, and Saturation (levels: color scenes, black-and-white scenes) manipulated between participants.

After exporting the recordings for each trial, we first annotated Speech Onset Times (SOT), manually. We defined SOT as the start of the utterance, excluding filled pauses, coughs, and sighs. The speech onset times served as the first dependent variable of the experiment. This variable was announced as an exploratory variable in our pre-registration at the OSF, and turned out to be of great value since it allowed us to replicate the analyses of domain size by Gatt et al. (2017) with photo-realistic scenes.

Before analyzing the eye tracking data, we defined one area interest (AOI) in every scene, which corresponded to the target object that was referred to in that specific scene. This AOI allowed us to have two dependent variables for the eye movement data. Firstly, we measured the proportion of target fixations: the number of times that the target AOI was fixated, relative to the total number of fixations in a specific scene. Secondly, we counted the number of switches in fixations back and forth between the target AOI and the rest of the scene. These were the two dependent variables that were announced in our pre-registration at the OSF.

Results

To analyze the data, we conducted Repeated Measures ANOVAs: one for all three dependent variables. Bonferroni tests

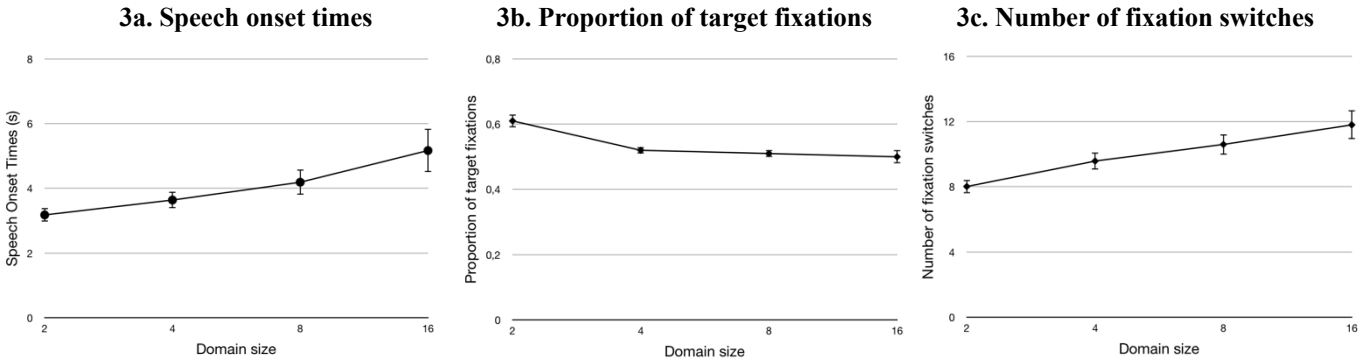

Fig. 3a-c: Means for speech onset times (3a), the proportion of target fixations (3b), and the number of fixation switches (3c) as a function of Domain size. Error bars represent ± standard error of the mean.

were applied for post hoc multiple comparisons. The data set consisted of 1056 referring expressions. In the analyses of the speech onset times, four trials were marked as missing values since the audio was not recorded properly. Data for two trials were missing in the eye tracking analyses, due to eye movements that failed to record.

Results for speech onset times

For SOTs, the ANOVA showed a main effect of Domain size ($F_{(3,126)} = 12.24$, $p < .001$, $\eta_p^2 = .23$). As can be seen in Figure 3a, the time until speech onset (in seconds) increased linearly as the number of distractors in the scene became higher. The post hoc tests revealed that the differences between scenes with two ($M = 3.18$, $SE = .21$), four ($M = 3.64$, $SE = .25$), and eight ($M = 4.19$, $SE = .39$) distractors were all significant (p always $< .05$), while the difference between scenes with eight and sixteen ($M = 5.17$, $SE = .67$) distractors was trending towards significance ($p = .054$).

The main effect of Saturation ($F_{(1,42)} = .00$, $p = .996$, $\eta_p^2 = .00$) was not found: the mean SOTs for the black-and-white ($M = 4.06$, $SE = .51$) and color ($M = 4.04$, $SE = .51$) conditions were practically indistinguishable. Also the interaction effect between Saturation and Domain size failed to reach significance ($F_{(3,126)} = .003$, $p = 1.00$, $\eta_p^2 = .00$), meaning that the pattern for Domain size was similar in both black-and-white and color scenes.

Results for the proportion of target fixations

The second ANOVA was run to analyze if our manipulations affected the number of times that participants fixated the target referent in a scene, relative to the total number of fixations in the trial. For Domain size, this was indeed the case ($F_{(3,126)} = 32.80$, $p < .001$, $\eta_p^2 = .44$). As can be seen in Figure 3b, this main effect was due to a higher proportion of target fixations in scenes with only two distractor objects ($M = .61$, $SE = .02$), as compared to scenes with four ($M = .52$, $SE = .01$), eight ($M = .51$, $SE = .01$), and ($M = .50$, $SE = .02$) distractors. The post hoc tests indeed revealed significant differences between the condition with the smallest domain size on the one hand, and the remaining three conditions on the other hand (p always $< .001$). The three biggest domain sizes resulted in similar proportions of target fixations (p always $> .71$).

Again, the main effect of Saturation ($F_{(1,42)} = 1.09$, $p = .30$, $\eta_p^2 = .03$) and the interaction between Saturation and Domain Size ($F_{(3,126)} = 1.45$, $p = .23$, $\eta_p^2 = .03$) were both not significant. For Saturation, black-and-white ($M = .52$, $SE = .02$) and color ($M = .55$, $SE = .02$) scenes led to similar proportions of target fixations, while the non-significant interaction shows implies that the patterns for the effect of Domain size were again the same in the two Saturation conditions.

Results for the number of fixation switches

The final ANOVA tested if the number of fixation switches back and forth between the target AOI and the rest of the visual scene was affected by Domain size and Saturation, again relative to the total number of fixations in the scene. Once more, the data revealed an effect of Domain size here ($F_{(3,126)} = 21.93$, $p < .001$, $\eta_p^2 = .34$), now showing a linear increase in the number of fixation switches in scenes with either two ($M = 8.02$, $SE = .40$), four ($M = 9.58$, $SE = .53$), and eight ($M = 10.6$, $SE = .62$) distractors. The post hoc tests showed that the differences in fixation switches between these conditions were indeed all significant (p always $< .03$). Although in the numerical sense, the number of fixation switches increased even further for big domains of sixteen distractors ($M = 11.8$, $SE = .88$), the difference between the means for eight and sixteen distractors was not significant ($p = .18$). The overall pattern of means and SEs is visualized in Figure 3c.

As with the previous variables, the main effect of Saturation ($F_{(1,42)} = .08$, $p = .78$, $\eta_p^2 = .002$) and the interaction between Saturation and Domain Size ($F_{(3,126)} = .11$, $p = .96$, $\eta_p^2 = .003$) were both not significant. Black-and-white scenes ($M = 10.1$, $SE = .80$) and color scenes ($M = 9.82$, $SE = .80$) resulted in the same proportions of fixations switches, and to similar patterns for the effect of Domain size in the two Saturation conditions.

Preliminary analysis of object descriptions

On top of the results for the SOT and eye tracking variables reported so far, the next step would be to see how our manipulations of domain size and saturation affect the actual referring expressions that were produced by our speakers. In the current paper, we present a preliminary analysis of the length of the object descriptions (in number of words), as a function

of the two independent variables. As we discuss below in the final part of the Discussion, analyzing the object descriptions in combination with SOTs and inspection patterns can tell us more about how incremental scene perception and reference production evolve.

We counted the numbers of words used to describe a target, excluding filled pauses (transcribed as "eh"). We found the number of words to increase linearly as the number of distractors in the scene became higher, resulting in a main effect of Domain size ($F_{(3,126)} = 43.93$, $p < .001$, $\eta_p^2 = .51$). The main effect of Saturation ($F_{(1,42)} = .10$, $p = .75$, $\eta_p^2 = .002$), and the interaction between Domain size and Saturation ($F_{(3,126)} = .85$, $p = .47$, $\eta_p^2 = .02$) were both not significant.

The means for the main effect of Domain size showed that, on average, 9.3 words ($SE = .68$) were used in scenes with two distractors; 10.8 words in the case of four distractors ($SE = .74$); 13.3 words in the case of eight distractors ($SE = .80$); and 14.3 words in the case of sixteen distractors ($SE = .93$). The post hoc tests showed that the differences in number of words between the conditions were all significant (p always $< .001$), except for the difference between scenes with eight and sixteen distractors (p always $= .24$).

Discussion

The goal of this research was to investigate how domain size affects speech onset times and eye movements during definite reference production. We manipulated scenes with one target and either 2, 4, 8 or 16 distractors. These scenes were applied in black-and-white and in color, to test if domain size interacts with saturation. The relevance of our study was twofold: we manipulated domain size in photo-realistic scenes rather than artificial grids of objects; and we collected participants' eye movements to take a direct measure of scene perception as a function of domain size.

Our manipulation of domain size was inspired by Gatt et al. (2017), who manipulated this variable in artificial grids of objects, and found a linear increase in onset times as domains got bigger, but only when the target did not "pop out" of the scene by means of its color (echoing earlier work in the vision literature; Treisman & Gelade, 1980). If we turn to the speech onset times observed in the current experiment, with photo-realistic scenes, we see a pattern that is similar to the one in Gatt et al (2017): also here, onset times increase linearly with bigger domains. For one thing, this suggests that the impact of pop out mechanisms was probably limited in our stimuli, but more importantly, it implies that our speakers performed object-by-object scans when processing the domains. Our eye tracking data provide evidence for this suggestion. Firstly, if we look at the proportion of target fixations, we see that target referents are fixated less frequently in bigger domains of 4, 8 and 16 distractors (rather than 2), which suggests that speakers in those cases put more effort in scanning the distractors, which are higher in number. Secondly, for the number of fixation switches, we see a linear increase as domains get bigger, showing that speakers are actually comparing targets against their distractors, and that they switch back and forth more often simply because there are more distractors that are relevant to consider.

Thus, based on our results, we argue that the photo-realistic nature of our stimuli did not prevent speakers from performing a careful scan of the objects in the scene. As we have seen, the vision literature shows that target search can get more efficient as scenes contain more distractors, because these extra objects provide context that allows viewers to rely on a set of only functional distractors (Neider & Zelinsky, 2008). This is not what our results seem to suggest. However, since our eye tracking analyses do not distinguish between individual distractors and their characteristics, they do not strictly rule out that certain distractors may be considered more relevant than others, and were therefore fixated more often. Further analyses of the data could shed more light on this matter.

In further analyses of the current data set, we are planning to also involve characteristics of the actual referring expressions in our analyses. As a first step, the current paper tested how our manipulations of domain size and saturation affected the number of words used by our speakers, in a preliminary analysis. The results are promising: we found a linear increase in the number of words for bigger domain sizes, similar to the increase in SOTs and fixation switches. This pattern reveals an interesting picture of how visually grounded reference production evolves incrementally: in the case of more distractors, speakers need more time to 'plan' an expression (i.e., longer SOTs), arguably because both visual processing (i.e., more fixation switches) and content planning (i.e., more words) become more elaborate.

In addition to the number of words, we believe that it would be even more relevant to also annotate and analyze the attributes that were mentioned by our speakers. As announced in our pre-registration at the OSF, our plan is to include Attribute type as an extra variable to the design, and to compare all descriptions that contain only absolute attributes (e.g., hair color) to all descriptions that include only relative attributes (e.g., age). This variable could affect our dependent variables, since the presence of relative attributes suggests that the target object has been compared to at least one of the distractors. Hence, it might also interact with our manipulation of Saturation, since one would expect more relative attributes to be mentioned in black-and-white scenes, where pop-out effects are less likely to occur than in color scenes.

Acknowledgments

We thank Rein Cozijn for his generous help in setting up the experiment, and analyzing the eye tracking data.

References

Belke, E. (2006). Visual determinants of preferred adjective order. *Visual Cognition, 14* (3), 261-294.

Davies, C. & Kreysa, H. (2017). Looking at a contrast object before speaking boosts referential informativity, but is not essential. *Acta Psychologica, 178*, 87–99.

Elsner, M., Clarke, A., & Rohde, H. (2018). Visual complexity and its effects on referring expression generation. *Cognitive Science, 42* (4), 940–973.

Gatt, A., Krahmer, E., Van Deemter, K., & Van Gompel, R. (2017). Reference production as search: the impact of domain size on the production of distinguishing descriptions. *Cognitive Science, 41* (6), 1457-1492.

Henderson, J., Brockmole, J., Castelhano, M., & Mack, M. (2007). Visual saliency does not account for eye movements during visual search in real world scenes. In R. van Gompel, M. Fischer, W. Murray, & R. L. Hill, (Eds.), *Eye movements: A window on mind and brain* (pp. 1–41).

Henderson, J. & Hollingworth, A. (1998). Eye movements during scene viewing: an overview. In G. Underwood (Ed.), *Eye guidance in reading and scene perception* (pp. 269-293.

Henderson, J., Malcolm, G., & Schandl, C. (2009). Searching in the dark: cognitive relevance drives attention in real-world scenes. *Psychonomic Bulletin & Review, 16* (5), 850-856.

Koolen, R., Goudbeek, M., & Krahmer, E. (2013). The effect of scene variation on the redundant use of color. *Cognitive Science, 37* (2), 395-411.

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

Neider, B. & Zelinsky, G. (2008). Exploring set size effects in scenes: identifying the objects of search. *Visual Cognition, 16* (1), 1-10.

Nordfang, M. & Wolfe, J. (2014). Guided search for triple conjunctions. *Attention, perception & psychophysics, 76* (6), 1535-1559.

Palmer, J. (1995). Attention in visual search: distinguishing four causes of a set size effect. *Current directions in psychological science, 4* (4), 118-123.

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

Rubio-Fernández, P. (2016). How redundant are redundant color adjectives? An efficiency-based analysis of color overspecification. *Frontiers in Psychology, 7*: 153.

Tanenhaus, M., Spivey, M., Eberhard, K., & Sedivy, J. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*, 1632-1634.

Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review, 113* (4), 766–786.

Treisman, A. & Gelade, G. (1980). A feature integration theory of attention. *Cognitive Psychology, 12*, 97-136.

Wolfe, J. (2010). Visual search. *Current Biology, 20* (8), R346–R349.