Eye movements during reference production: Testing the effects of perceptual grouping on referential overspecification

Ruud Koolen (r.m.f.koolen@uvt.nl)
Tilburg center for Cognition and Communication (TiCC), Tilburg University
PO Box 90153, 5000 LE, The Netherlands

Yorick Fliervoet (yorick@fliervoet.com)
Tilburg center for Cognition and Communication (TiCC), Tilburg University
PO Box 90153, 5000 LE, The Netherlands

Abstract

When referring to a target object in a visual scene, speakers are assumed to consider certain distractor objects that are visible to be more relevant than others. However, previous research that has tested this assumption has mainly applied offline measures of visual attention, such as the occurrence of overspecification in speakers' target descriptions. Therefore, in the current study, we take both online (eye-tracking) and offline (overspecification) measures of attention, to study how perceptual grouping affects scene perception, and reference production. We manipulated three grouping principles: region of space, type similarity, and color similarity. For all three factors, we found effects, either on eye movements (region of space), overspecification (color similarity), or both (type similarity). The results for type similarity provide direct evidence for the close link between scene perception and reference production.

Keywords: Reference production; Perceptual grouping; Eye-movements; Overspecification; Visual scene perception.

Introduction

Suppose you want to point out the marked object in Fig. 1 to a listener. To complete this task, you should produce a referring expression such as "the small bowl" or "the small green bowl", to distinguish the *target object* from the other objects that are present in the visual scene (the *distractors*). Although both above example expressions allow the listener to identify the target, the second one is *overspecified*: it contains a color attribute that is unnecessary for unique identification.

Figure 1: An example visual scene (Koolen et al., 2014)

From prior research (e.g., Pechmann, 1989; Koolen, Goudbeek, & Krahmer, 2013; Rubio-Fernàndez, 2016), it is known that speakers overspecify their referring expressions very frequently. Why do they do so? We argue that at least one of the answers to this question is to be found in *visual scene perception*, and explore to what extent certain objects in a scene are more likely to be perceived than others. For example, in Fig. 1, the plate on the sideboard might be overlooked because it is placed on a different surface than the target (i.e., sideboard rather than table), or because it has a different type (i.e., plate rather than bowl). In these cases, the distractor set would be limited to the large bowl, making a minimal description such as "the small bowl" likely to be uttered. On the other hand, if the plate on the sideboard catches attention anyway, for example because it has a different color than the target object, the perceived color variation may cause speakers to overspecify with color (Koolen et al., 2013).

Although there is growing awareness that scene perception and language production are indeed closely linked, previous research in this direction has generally taken indirect, offline measures of visual attention. Therefore, in the current paper, we combine online (eye-tracking) and offline (occurrence of overspecification) measures to search for structural relations between scene perception and attribute selection for referring expressions.

Theoretical background

The starting point of our research is the assumption that in a reference production task, speakers do not regard all objects in a visual scene to be relevant distractors, but rather rely on a subset of distractor objects. More specifically, speakers are expected to only consider the distractors that are in their focus of attention (Beun & Cremers, 1998). One can think of various factors that determine whether an object is perceived or not, such as its physical distance to the target (i.e., proximity). Given that proximity predicts that only objects that are close to the target referent are in the speaker's focus of attention, it can influence the composition of the distractor set for a visual scene (Clarke, Elsner, & Rohde, 2013a).

Proximity is one of the Gestalt laws of *perceptual grouping* that were originally introduced by Wertheimer (1923), next to similarity, closure, continuation, and pragnanz. These laws are principles of perceptual organization that serve as heuristics: mental shortcuts for how we perceive the visual environment (Wagemans et al., 2012) and create meaningful groups of objects that we see around us (Thórisson, 1996). On top of the classical laws of grouping, Palmer (1992) defined another principle, common region of space, which holds that objects

that fall within an enclosing contour, such as a table surface, are usually perceived as a group as well.

This study will apply a manipulation of common region of space, as well as two manipulations of similarity: color similarity and type similarity. Previous research that directly tests how these principles influence reference production is scarce. For color similarity, we know that speakers overspecify more often when they perceive color variation in a scene than when all objects are of the same color (Koolen et al., 2013; Rubio-Fernàndez, 2016). This effect of color variation interacts with type similarity: the proportion of overspecification is highest when there is at least one distractor object that shares its type with the target, but not its color (Koolen, Krahmer, & Swerts, 2016). Also common region of space has been found to affect referential overspecification, as revealed by Koolen, Houben, Huntjens, and Krahmer (2014). In their experiment, Koolen et al. used scenes such as the one depicted in Fig. 1, displayed in both 2D and 3D. The target was always on the table, and – mainly for the 3D scenes – speakers overspecified more often when a differently colored distractor was also on the table (in the same group as the target) rather than on the sideboard (in a different group), although the physical distance between the objects was the same in both scenarios.

Crucially, the above papers, as well as many others studies on reference production (e.g., Clarke et al., 2013a), have used indirect measures of visual attention, such as the occurrence of overspecification. This is problematic in studying how the distractors in a visual scene shape attribute selection. For example, although the experiment by Koolen et al. (2014) suggests that region of space affects overspecification, there is no direct evidence that this result is due to the way in which speakers might ignore distractors that are not in the same region as the target referent. Therefore, in the current research, we collect eye movements as a direct, online measure of visual attention, and combine these data with a more traditional, offline analysis of referential overspecification.

While eye-tracking methodologies are very commonly used to investigate language comprehension (e.g., Tanenhaus, Spivey, Eberhard, & Sedivy, 1995), they are still rare in language production research, initially because speech movements can disrupt eye movement data (Pechmann, 1989; Griffin & Davison, 2011). After some early studies that explored the effect of object fixations on order of mention (e.g., Griffin & Bock, 2000; Meyer, Sleiderink, & Levelt, 1998), some researchers recently started to apply eye-tracking to test the effects of perceptual and conceptual scene properties on rather open-ended descriptions (Coco & Keller, 2012; 2015) and object naming (Clarke, Coco & Keller, 2013b). However, none of this work has tested systematically how perceptual grouping affects attribute selection for reference production.

Current study

To study how different manipulations of perceptual grouping affect reference production, we conducted an experiment in which speakers described target objects in visual scenes. The stimuli were taken from Koolen et al. (2014), for the sake of comparability. We recorded both the participants' speech as well as their eye movements during the reference production task. Speech data were annotated for the occurrence of overspecification; i.e., if descriptions contained a redundant color attribute. This variable served as a replication of Koolen et al. (2014). New in our study are the eye-tracking data. Here, we analyzed the number of fixations on the distractor we manipulated, and the total gaze duration for that object.

For region of space, we hypothesize that if a distractor is in the same region of space as the target, it is viewed more often and longer than if the region of space is different, and that this will eventually lead to more overspecification. The same goes for type similarity, with more views, longer viewing time and more overspecification for a distractor of the same rather than a different type than the target. Lastly, for color similarity, we expect to find that a distractor most likely attracts attention if it has a different color than the target, resulting in more views, longer viewing times, and again more overspecification than for a distractor that shares its color with the target.

Method

Participants

Thirty-one participants (26 female, mean age: 21.6) took part in the experiment. The participants were gathered randomly at the campus of Tilburg University, and received a piece of candy as a reward. All participants were native speakers of Dutch, the language of the experiment.

Materials

The stimulus material consisted of near-photorealistic visual scenes like the example scenes presented in Fig. 2 on the next page. As noted above, the scenes were taken from the related previous study by Koolen et al. (2014). They depicted a living room containing a dinner table and a sideboard, and some objects such as chairs for a more realistic look. The scenes were modeled and rendered using Maxon's Cinema 4D.

The table and the sideboard formed the two surfaces (i.e., regions of space) that were important for our manipulations, since these were the spaces where the target and its two distractors were positioned. The target object always occurred on the table, in the middle of the scene, together with a distractor close next to it (either left or right). This first distractor object always had the same type and color as the target object, but a different size. This way, the distractor ensured that size was always needed for a distinguishing description, and that mentioning color thus resulted in an overspecified referring expression. The scenes also had a second distractor object, by means of which our three manipulations of perceptual grouping were realized.

Firstly, there was a manipulation of perceptual grouping in the law of *common region of space*. This manipulation was operationalized by positioning the second distractor either in the *same region* of space as the target (i.e., on the table, see the left scenes of Fig. 2), or in a *different region* (i.e., on the sideboard, see the right pictures of Fig. 2). It is important to note that the physical distance between the target object and the second distractor was the same in both scenarios.

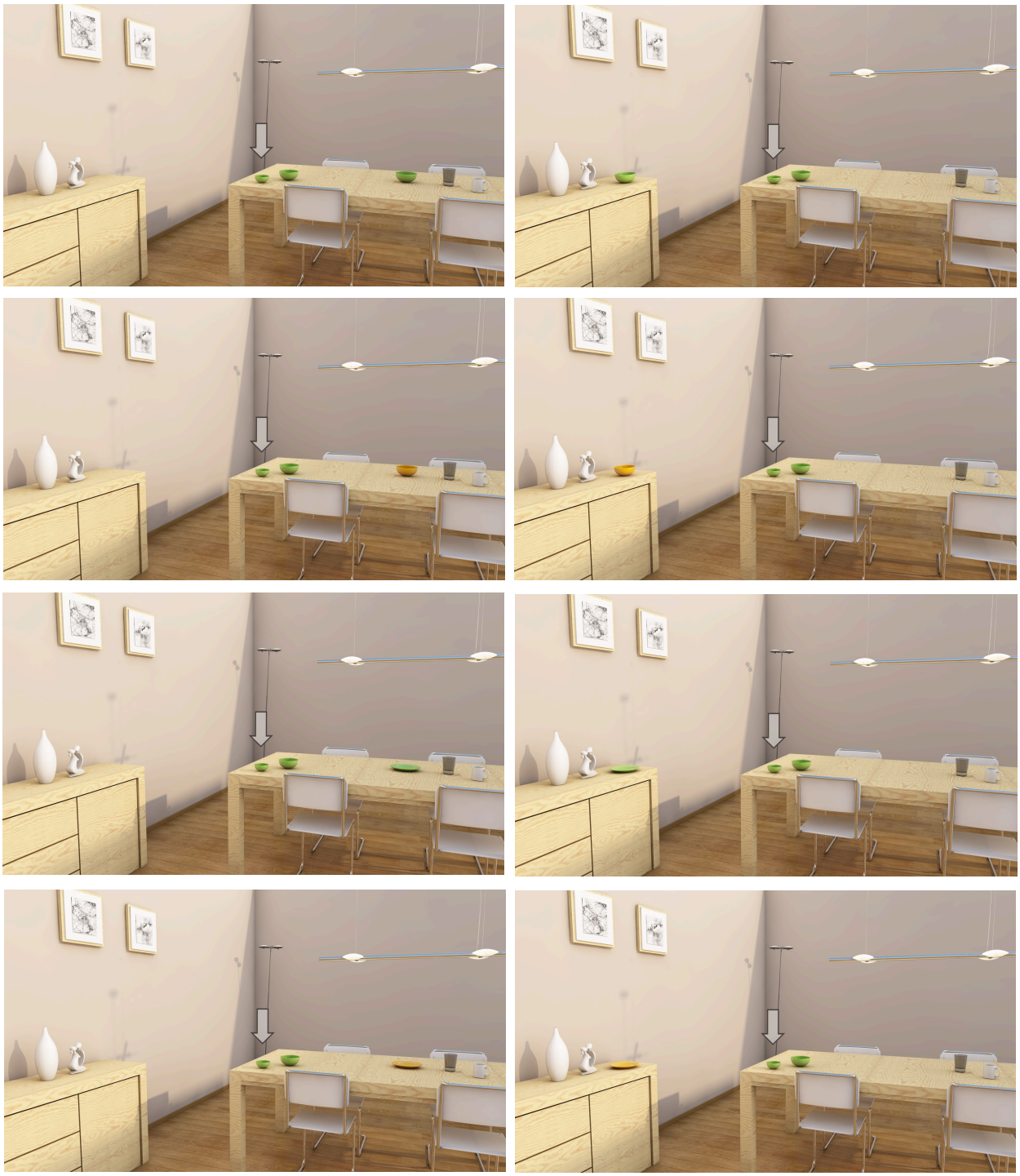

Fig. 2: Examples of critical trials in our experiment. The distractor shares its region of space with the target (i.e., the table) in the left scenes, and is in a different region (i.e., the sideboard) in the right scenes. The distractor has the same type as the target in the upper four pictures, and a different type in the lower four pictures. The distractor has the same color as the target in the first, second, fifth and sixth picture, and a different color in the third, fourth, seventh, and eighth picture.

Secondly, we had two manipulations of perceptual grouping related to the law of similarity. The first one varied the *type* of the distractor: this type could be *the same* as the target's type, or *different*. Example scenes can again be found in Fig. 2, where the second distractor object (the plate) has a different type than the target object (the bowl) in the lower four trials, while all relevant objects are of the same type in the upper four trials. Another manipulation of similarity was

employed by varying the *color* of the second distractor: this color could again be *the same as* or *different than* the color of the target object (see again Fig. 2 for example scenes).

While Fig. 2 depicts all visual scenes that were created for the bowl, the same was done for three other types of targets: a plate, a mug and a cutting board. The scenes for these four object types were manipulated in all conditions, resulting in eight trials for each object type. Participants were thus presented with thirty-two (four x eight) critical trails. In all trials, the target object could only be distinguished by mentioning type and size; if participants included color, it made the description overspecified.

Two measures were taken to avoid participants from using the same strategy for all critical trials. Firstly, we had thirty-two filler trials. Although the scenes for these fillers had the same basic set-up as the critical trails, with all kinds of objects placed on the table and the sideboard, there were more objects present, which could all be the target for that scene. Furthermore, since all objects in the filler trials were white, participants were discouraged to use color when referring to the target here.

Secondly, to prevent participants from developing a viewing strategy, we created two versions of the experiment. For both versions, half of the visual scenes for the critical trials were mirrored. In version 1, this was done for the scenes in which the second distractor was in the same region of space as the target object, while in version 2, all the scenes in the different region of space condition were mirrored. Thus, by taking this measure, all participants saw half of the critical trails mirrored.

Procedure

The experiment took place in a soundproof booth, located in the SensoMotoric Instruments lab at Tilburg University. The eye-tracking measurements were made with a SMI RED250 device, operated by the IviewX and the ExperimentCenter software-packages. The eye-tracker had a sampling rate of 250HZ. We used the microphone of a webcam to record the descriptions of the participants; the camera was taped off for privacy reasons. The stimulus materials were displayed on a 22 inch P2210 Dell monitor, with the resolution set to 1680x 1050 pixels, with 90.05 pixels per square inch.

After entering the laboratory, participants signed a consent form, and read a first basic instruction stating that they were going to act as the speaker in a language production experiment. Participants were then seated in the soundproof booth, in front of the eye tracker, and their eyes were calibrated using a 9-point validation method. When the calibration was completed successfully, participants were invited to read a second instruction, which was more detailed than the first one, and stated that participants were going to produce oral descriptions of target objects in visual scenes in such a way that these objects could be distinguished from the remaining objects in the scene. It was emphasized that using location information in the descriptions (e.g., "the bowl on the left") was not allowed. After this second instruction, participants completed two practice trials, and had the possibility to ask questions. Once the procedure was clear, the experimenter left the booth, and the experiment started.

All participants were shown a total of 64 stimuli (32 critical trails and 32 fillers) in a random order. The visual scenes were depicted in the middle of the screen, filling 70% of the available space; the remaining 30% consisted of a grey border surrounding the scenes. Before every trial, a screen with an 'X' appeared somewhere in the 30% contour area. When this X had been fixated for one second, the next visual scene appeared automatically. When fixating the X did not work, participants could make the next scene appear manually by pressing spacebar. The position of the X was different for all trials: they appeared in a random position in the grey border, again to make sure that participants did not develop a viewing strategy. There were 1.6 times more X triggers on the top and bottom row than on the left and right side, in proportion to the 1680x1050 screen resolution. Once all 64 trials had been completed, participants were instructed to leave the booth. It took around 30 minutes to complete the experiment.

Research design

The experiment had a 2x2x2 design with three within-participants factors: *region of space* (same, different), *type* (same, different), and *color* (same, different). Three dependent variables were measured: the occurrence of color in the target descriptions; the gaze duration upon the manipulated distractor in milliseconds per trial per participant; and the number of times that the manipulated distractor was fixated per trial per participant.

Data coding and preparation for analysis

All recorded object descriptions were transcribed and coded for the presence of color (0 or 1). For the eye-tracking data, we first checked for ill measurements, and excluded the data recorded for one participant from further analysis. We then assigned all fixations to either one out of four areas of interest (AOIs) we defined. There was one AOI for the target, one for the sideboard, one for the central part of the table, and one remainder area. The AOIs for the sideboard and the central part of the table represented the areas where the manipulated distractor could be placed. The remainder area was used for fixations that were not on the target or distractor objects that were present in the scenes. The AOIs where the manipulated distractor could occur were central to our analyses.

The coding process resulted in a separate path file for every participant. These path files were converted into a single file, and loaded into SPSS for statistical analysis. Although there was supposed to be data for 960 target descriptions (30 speakers times 32 trials), the data for 24 trials could not be analyzed because either the description or the eye movements were not recorded correctly. The final analysis thus contained data for 936 trials.

While the data for all 936 trials was used to analyze the redundant use of color, we created subsets of the data to analyze gaze duration and the number of fixations. For both variables, we only analyzed the cases where speakers fixated – and thus saw – the manipulated distractor. This was the case in 680 out

of 936 cases. For gaze duration, we then calculated for every trial the total amount of time that the participant looked at the manipulated distractor object, and standardized this score by calculating the *z*-score per trial per speaker. Only the scores in the range of $-3 \leq z \leq 3$ were included in the analysis, which means that scores for 13 cases were filtered out.

For the number of fixations, we created a similar subset of the data, but this time we calculated the number of times that speakers looked at the manipulated distractor for every trial. Again, the *z*-score was calculated, which led to the exclusion of 12 trials that were not part of the final analysis for this variable.

Results

To test for significance, we performed a series of univariate ANOVA tests. We only report on interactions when they are significant. Given that we used subsets of the data in our statistical analyses, performing repeated measures tests was not possible due to empty cells.

Results for redundant color use

In general, our speakers included a redundant color attribute in 64% of the descriptions. The first ANOVA was performed to test if redundant color use was affected by our manipulations of perceptual grouping.

The first factor that we expected to affect the redundant use of color was region of space. However, we did not find a significant effect here ($F_{(1,927)} = .11$, n.s.): speakers redundantly used color equally often when the manipulated distractor was in the same (M = .64, SE = .02) or a different (M = .64, SE = .02) region of space as compared to the target.

For our two manipulations of similarity, we did find effects on the redundant use of color. In these cases, the main effects of type similarity ($F_{(1,927)} = 9.94$, $p < .01$, $\eta_p^2 = .011$) and color similarity ($F_{(1,927)} = 5.44$, $p < .05$, $\eta_p^2 = .006$) were due to an increase in redundant color use when the manipulated distractor had the same type as the target, and a different color (M = .77, SE = .03). The other three cells were practically indistinguishable (same type - same color: M = .61, SE = .03; different type - same color: M = .60, SE = .03; different type - different color: M = .59, SE = .03). This pattern resulted in a significant interaction between type similarity and color similarity ($F_{(1,927)} = 7.47$, $p < .01$, $\eta_p^2 = .008$).

Results for gaze duration

The second ANOVA was run to analyze if our manipulations of grouping on the total amount of time that speakers looked at the manipulated distractor.

Firstly, there was a main effect of region of space on gaze duration ($F_{(1,651)} = 215.5$, $p < .001$, $\eta_p^2 = .249$), showing that the distractor object was looked at significantly longer when it occurred in the same (M = 1812.7, SD = 60.87) rather than a different (M = 466.7, SE = 68.6) region of space than the target. A similar effect was found for the manipulation of type similarity ($F_{(1,651)} = 5.06$, $p < .05$, $\eta_p^2 = .008$). For this factor, we found that distractors that shared their type with the target (M = 1242.9, SE = 66.2) were looked at longer than distractors for which this was not the case (M = 1036.5, SE = 63.5). The third factor, color similarity, did not affect gaze duration: although the distractor was looked at slightly longer when it had the same (M = 1176.6, SE = 61.6) rather than a different (M = 1102.8, SE = 67.9) color than the target, this difference was not significant ($F_{(1,651)} = .65$, n.s.).

Results for number of fixations

The third dependent variable in our experiment was the number of fixations on the manipulated distractor. Again, there were effects of region of space and type similarity, but not of color similarity.

Firstly, when the distractor was in the same region of space as the target object, participants looked at this object significantly more often (M = 2.04, SD = .06) than when it occurred in a different region of space (M = 1.56, SD = .06); $F_{(1,652)} = 33.37$, $p < .001$, $\eta_p^2 = .049$. Similarly, when the distractor was of the same type as the target object, it was fixated more often (M = 1.93, SD = .06) than when it had a different type (M = 1.67, SD = .06). Again, we found no effect of color similarity: the distractor's color (same: M = 1.85, SE = .06; different: M = 1.76, SE = .06) did not influence the number of fixations ($F_{(1,652)} = 1.15$, n.s.).

Discussion

The goal of this research was to test how perceptual grouping affects reference production. We combined both online (eye-tracking) and offline (occurrence of referential overspecification) measures of visual attention to study the extent to which grouping causes speakers to ignore certain distractors that are present in a visual scene, aiming to connect the observed scan patterns referential overspecification. We had three manipulations of grouping (i.e., common region of space, color similarity, and type similarity), all realized by varying the location and characteristics of one specific distractor object in the visual scenes that were presented to the participants.

The first manipulation that was present in our stimuli made the manipulated distractor object appear either in the same or a different *region of space* as compared to the target referent. In Koolen et al. (2014), this manipulation led to a significant effect of grouping on overspecification, with more redundant color attributes in the 'same group' condition rather than the 'different group' condition. In the current study, we could not replicate this result: the proportions of overspecification that we found were the same in both conditions. However, we did find effects of region of space in the eye-tracking data: when the distractor was in the same region as the target referent, it was viewed longer and more often than when it was in a different region. This way, region of space (Palmer, 1992) influences the extent to which certain distractors are considered in a reference production task.

The question remains why the patterns for common region of space that we observed in the eye-tracking data were not reflected in effects on overspecification with color, such as found by Koolen et al. (2014). To explain this issue, we refer to some practical differences between the two studies. Firstly,

Koolen et al. (2014) displayed the stimuli on a big television screen, while the current experiment used only 70% of a computer screen. Perhaps more important was that Koolen et al. found a convincing effect of common region of space for 3D visual scenes, but that the effect was small for 2D scenes. In the current study, only 2D scenes were used, due to the eye-tracking paradigm. Given that our 2D scenes led to clear effects of region of space in the eye-tracking data, it would be interesting to test how this grouping principle affects language on variables other than overspecification, such as fluency and speech onset time.

For *type similarity*, the effect of the manipulation in the reference production data resonates the pattern in the eye-tracking data. When the distractor had the same type as the target, it was viewed longer and more often than when the type was different, and the proportion of overspecified references was higher. These results show direct evidence for the close link between visual scene perception and language production, in line with the few previous studies in this direction (e.g., Coco & Keller, 2012; 2015; Griffin & Bock, 2000). For *color similarity*, we found a significant interaction with type similarity for the speech data, with an increase in overspecification with color when the distractor had the same type as the target, and a different color. This interaction is a replication of Koolen et al. 2016). For the eye-tracking data, there were no significant effects or interactions with color similarity involved, presumably since color differences "pop out" of the scene (Treisman & Gelade, 1980). As such, there is no strict need for speakers to fixate distractors (repeatedly) in order to perceive their different color.

Finally, we would like to discuss our decision to use subsets of the data for the eye-tracking analyses. In these subsets, we only included data for the trials where the speaker fixated the manipulated distractor object (or at least the AOI where it was occurred) at least once. Thanks to this approach, we could be certain that speakers were most likely aware of the existence of this object, which makes the observed effects of perceptual grouping even more valid: it excludes, for example, measurement errors that occur when speakers change their position in front of the eye-tracker. However, one can also argue that our approach was too strict, because in order to form a description of a target object, it is not necessary to scan all objects in the scene. In future analyses, we aim to refine our paradigm, also by distinguishing various time windows for every trial to test both the structural and temporal relations between scene perception and reference production.

Acknowledgments

We thank Jan Huntjens and Eugène Houben (www.eyetractive.com) for developing the stimulus materials, and Rein Cozijn for his assistance in programming the experiment and analyzing the data.

References

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

Clarke, A., Elsner, M., & Rohde, H. (2013a). Where's Wally: the influence of visual salience on referring expression generation. *Frontiers in Psychology, 4*, article 329.

Clarke, A., Coco, M. & Keller, F. (2013b). The impact of attentional, linguistic, and visual features during object naming. *Frontiers in Psychology, 4*: article 927.

Coco, M. & Keller, F. (2012). Scan patterns predict sentence production in the cross-modal processing of visual scenes. *Cognitive Science*, *36* (7), 1207-1223.

Coco, M. & Keller, F. (2015). Integrating mechanisms of visual guidance in naturalistic language production. *Cognitive Processing 16* (2), 131-150.

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

Griffin, Z. & Davison, J. (2011). A technical introduction to using speakers' eye movements to study language. *The Mental Lexicon*, *6* (1), 53-82.

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

Koolen, R., Houben, E., Huntjens, J., & Krahmer, E. (2014). How perceived distractor distance influences reference production: Effects of perceptual grouping in 2D and 3D scenes. In *Proceedings of the 36th annual meeting of the Cognitive Science Society (CogSci)*. Québec, Canada.

Koolen, R., Krahmer, E., & Swerts, M. (2016). How distractor objects trigger referential overspecification: testing the effects of visual clutter and distance. *Cognitive Science, 40* (7), 1607-1647.

Meyer, A., Sleiderink, A. & Levelt, W. (1998). Viewing and naming objects: eye movements during noun phrase production. *Cognition, 66*, B26-B33.

Palmer, S. (1992). Common region: a new principle of perceptual grouping. *Cognitive Psychology, 24* (3), 436-447.

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.

Wagemans, J., Elder, J., Kubovy, M., Palmer, S., Peterson, M., Singh, M., & Von der Heydt, R. (2012). A century of Gestalt psychology in visual perception: I. Perceptual grouping and Figure–ground organization. *Psychological Bulletin, 138* (6), 1172.

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

Thórisson, K. (1994). Simulated perceptual grouping: an application to human-computer interaction. *Proceedings of the 16th annual conference of the Cognitive Science Society (CogSci)*, 876-881. Atlanta, Georgia, USA.

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

Wertheimer, M. (1923). Untersuchungen zur Lehre von der Gestalt. *Psychologische Forschung*, 4, 301-350.