Viewing time affects overspecification: Evidence for two strategies of attribute selection during reference production

Ruud Koolen (r.m.f.koolen@uvt.nl)
Tilburg center for Cognition and Communication (TiCC), Tilburg University

Albert Gatt (albert.gatt@um.edu.mt)
Institute of Linguistics, University of Malta

Roger P.G. van Gompel (r.p.g.vangompel@dundee.ac.uk)
School of Psychology, University of Dundee

Emiel Krahmer (e.j.krahmer@uvt.nl)
Tilburg center for Cognition and Communication (TiCC), Tilburg University

Kees van Deemter (k.vdeemter@abdn.ac.uk)
Department of Computing Science, University of Aberdeen

Abstract

Speakers often produce definite referring expressions that are overspecified: they tend to include more attributes than necessary to distinguish the target referent. The current paper investigates how the occurrence of overspecification is affected by viewing time. We conducted an experiment in which speakers were asked to refer to target objects in visual domains. Half of the speakers had unlimited time to inspect the domains, while viewing time was limited (1000 ms) for the other half. The results reveal that limited viewing time induces the occurrence of overspecification. We conjecture that limited viewing time caused speakers to rely heavily on quick heuristics during attribute selection, which urge them to select attributes that are perceptually salient. In the case of unlimited inspection time, speakers seem to rely on a combination of heuristic and more deliberate selection strategies.

Keywords: Definite reference; overspecification; heuristics; viewing time.

Introduction

In everyday language use, speakers often refer to objects in the world around them. Such references often take the form of a definite description that contains an article, one or more modifiers, and a head noun (e.g., "the brown chair", or "the large table"). Among the various reasons why speakers refer to objects, identification is perhaps the most obvious one: in many cases, speakers aim to distinguish one *target* from the *distractors* that are present in the context. Therefore, deciding what to say (or, in the terminology of Levelt, 1989, conceptualization) is a crucial part of referring: which attributes should be selected to make the target identifiable?

For this attribute selection process, previous research has shown that speakers tend to *overspecify* their object descriptions: they often use redundant attributes that are not strictly needed for unique identification of the target. For example, speakers might produce "the green chair" in a visual domain where only one chair is present. In general, one can say that overspecification is most likely to occur when speakers refer to target objects in rather complex visual domains, including cluttered domains (e.g., Clarke, Elsner & Rohde, 2013), and domains with a high amount of visual variation between the target referent and its distractors (e.g., Koolen, Goudbeek & Krahmer, 2013; Rubio-Fernández, 2016).

So why do speakers overspecify their object references so frequently? The answer to this question may be found in the incremental nature of speech production (Pechmann, 1989), and, in particular, in the recent suggestion that speakers rely on quick *heuristics* during attribute selection. Heuristics can be defined as "beliefs concerning the likelihood of uncertain events (...) that reduce the complex tasks of assessing probabilities and predicting values to simpler judgmental operations" (Tversky & Kahneman, 1974, p. 1124). Recent work in psycholinguistics has suggested that heuristics also affect attribute selection during reference production. In particular, it has been argued that speakers are in some cases prevented from making exact calculations about the shortest possible description in a referential domain (Van Deemter, Gatt, Van Gompel & Krahmer, 2012).

If speakers indeed use heuristics during attribute selection, this could explain their tendency to overspecify. We want to argue that heuristics cause speakers to include attributes that are salient in a given domain. In many referential situations, speakers are simply lacking time and / or cognitive effort to perform a careful object-by-object scan of the target and its distractors. Instead, they might rely on a heuristic, and select attributes that are easily and quickly processed by the visual system. The most notable example here is color, which is an absolute attribute that does not require comparison to any other object in the domain in order to be perceived. In other words, color "pops out" of the scene (Belke & Meyer, 2002; Treisman & Gelade, 1980). As a result, color is a preferred attribute, which is included irrespectively of the number of distractors that it rules out in a domain (Dale & Viethen, 2009). Also size can be salient, when size differences between the objects in the domain are sufficiently large (Van

Gompel, Gatt, Krahmer & Van Deemter, 2014), or due to speakers' general tendency to repeatedly include the same attributes during reference production (Tarenskeen, Broersma & Geurts, 2015).

The use of heuristics may serve as a plausible explanation for the occurrence of overspecification: if speakers include attributes based on salience and preference, they may end up with a description that contains attributes that are not strictly needed to identify the target referent. Furthermore, in some situations, also less preferred attributes might be required to rule out some last remaining distractors in the domain. Thus, given that speakers and listeners cooperate during referential communication (Brennan & Clark, 1996), it seems plausible to assume that attribute selection is usually determined by a combination of preference and discriminatory power (i.e., the number of distractors that an attribute excludes in a particular domain). This assumption implies that there are multiple influences on attribute selection, which has been modeled in various recent computational interpretations of the conceptualization process. For example, the *Visible Objects Algorithm* (Mitchell, Van Deemter & Reiter, 2013) models these influences as two consecutive stages: an early stage in which visually salient and preferred attributes are selected, and a later stage where attributes are included based on their discriminatory power. Also the algorithm proposed by Gatt, Goudbeek and Krahmer (2011a) distinguishes between heuristic and more stable strategies, but models them in parallel rather than consecutively.

Despite the above computational models, prior research in psycholinguistics that has tested the procedures proposed in these models is mostly lacking. Therefore, in this paper, our goal is to investigate if speakers are indeed guided by both heuristic and more deliberate processes during attribute selection. In order to reach this goal, we performed a reference production experiment where we manipulated *viewing time*: half of our speakers could take as much time as necessary to inspect the visual domains they were presented with, while the other half had limited viewing time. To see how viewing time influenced reference production, we analyzed speakers' tendency to overspecify their descriptions. Our expectations are as follows.

For speakers with unlimited viewing time, we expect that they will overspecify their referring expressions, but only to some extent. For one thing, heuristic viewing strategies will cause them to include properties that are perceptually salient for them. However, since there is unlimited time to inspect the visual domain here, we expect selection based on inherent salience to interact with selection based on discriminatory power and exact calculations. As a result, the amount of overspecification may be small: preferred attributes may be avoided in some cases, and speakers might rather select attributes to exclude as many distractors at once as possible. This strategy allows the listener to rule out the remaining distractor objects and identify the target (Olson, 1970).

For speakers who have limited time to inspect a scene, the situation might be different: in this case, they may not have enough time to calculate the shortest possible description and to perform an object-by-object scan of the visual scene. Instead, these speakers may base attribute selection heavily on inherent preferences for certain attributes, and might thus be guided primarily by heuristics rather than discriminatory power. For example, they might start uttering a description before they have scanned all distractor objects that are present in the visual scene. As argued by Pechmann (1989), this incremental process may in turn cause speakers to overspecify, since it makes them include salient attributes for which they are not sure if they are needed for identification or not. After all, the limited viewing time could prevent them from taking the listener perspective into account (Horton & Keysar, 1996), and from searching for attributes with the highest discriminatory power.

Method

To study the effect of viewing time on overspecification, we performed an experiment in which participants took part in a simple director-matcher task. In this task, one participant – the speaker – described a target referent from a group of seven objects, to a listener who saw the same objects but in a different configuration. The speaker was instructed to refer to target objects in such a way that the listener could identify the intended referents.

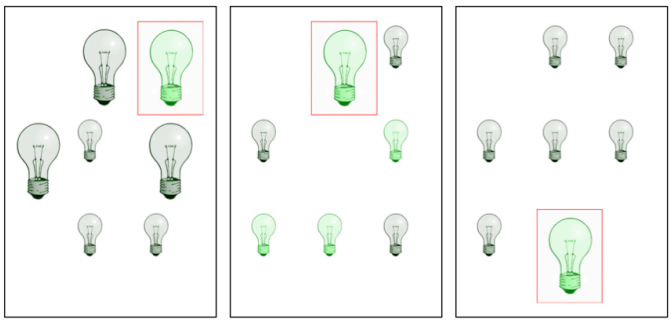
Participants

Participants were 36 undergraduate students from Tilburg University (19 female, 17 male, age range 17-34 years old, M = 21 years and 9 months), who took part as fulfillment of course credits. The participants were all native speakers of Dutch (the language of the study), and took part in the role of the speaker. A confederate – who was the same person for all participants – took part in the role of the listener.

Materials

The stimulus material consisted of thirty-six visual domains such as those shown in Fig. 1 on the next page. All domains depicted one target referent, which was marked with a red square, and six distractors. We constructed thirty-six domains that served as critical trials, and selected thirty-six different objects for the target objects. These objects were the same as those used in an earlier experiment by Gatt, Van Gompel, Krahmer, and Van Deemter (2011b). In all domains, the six distractors had the same type as the target referent. The seven objects were placed randomly in a 3x3 grid, such that the positions of the objects and the empty cells varied across domains.

Again following Gatt et al. (2011b), trials were manipulated in three *conditions*: (i) **C**, where *color* was required to identify the target referent; (ii) **S**, where *size* was necessary; and (iii) **C/S**, where participants could either use *color or size* to produce a distinguishing description. The use of size in the C condition, or color in the S condition, excluded only three out of six distractor objects. Example domains for the three conditions can be found in Fig. 1a-c. The thirty-six target objects were distributed across three lists, so that each



(a) Color needed (b) Size needed (c) Color or size needed

Figure 1: Experimental domains. The manipulations of color may not be visible in a black and white print of this paper.

target referent appeared in a different condition in each list. Participants were randomly assigned to one of the three lists, and the conditions always had twelve trials each.

Color contrast between the target and the distractors in the domains was manipulated in such a way that speakers could not use basic color terms to refer to a target's color. In Figure 1a-c, for example, speakers naturally used *light green* when their description included color, because some (or all) of the distractor objects were dark green. In half of the critical trials, the target's color shade was light, and in the other half it was dark. Target objects occurred either in red, grey, blue, or green.

The use of basic color terms was disabled in an attempt to balance the trade-off between speaker preferences for color and size attributes. As we have seen in the Introduction, it is known that speakers prefer to use color, even if this leads to overspecification (e.g., Pechmann, 1989; Belke & Meyer, 2002; Koolen et al., 2013, among many others). By contrast, attributes such as size tend to be used only when absolutely required. However, the inclusion of color becomes less likely when the color differences between the target and its distractors are small, and when basic color terms are not sufficient to identify the target (Viethen, Goudbeek & Krahmer, 2012). Both these conditions were met in the current experiment.

The second independent variable of the experiment, *pace*, manipulated viewing time. This variable was tested between participants. Half of the speakers took part in the *self-paced* condition, and could take as much time as needed for each trial to inspect the domain and describe the target. The other half of the speakers took part in the *system-paced* condition, Although the speakers in this condition could again take as much time as they needed to describe the target, the visual domains disappeared automatically after 1000 milliseconds for each trial. This means that the time that speakers had to inspect the domains and to find the distinguishing attributes for the targets was limited. The time window of 1000 milliseconds was decided upon with a pre-test. The main criterion here was that participants should experience pressure, but should still be given enough time to take a look at all objects in the domain, and thus to avoid underspecified descriptions that do not contain enough information to identify the target. From the pre-test, we learned that speakers were able to avoid underspecification when they were given 1000 milliseconds to inspect the domain; speakers confirmed that they could then indeed to take a look at all the objects that were visible. In the experiment itself, we found that only 1.3% of the descriptions were underspecified. The visual scenes for the confederate listener were always displayed for the duration of the whole trial, irrespective of condition.

In addition to the thirty-six domains for the critical trials, we created thirty-six filler domains. The fillers consisted of two abstract 3D 'Greebles' figures (Gauthier & Tarr, 1997), all purple, so that speakers were not primed with using color in the critical trials. One Greeble was marked as the target, and could be distinguished from the other Greeble by means of its main shape or by the direction in which its protrusions were pointing.

Procedure

The experiment took place in a quiet office room at Tilburg University. The average running time was approximately 15 minutes for each participant. After signing the consent form, participants were randomly assigned to either the self- or the system-paced condition, which resulted in eighteen speakers per condition. Participants sat at a table facing their listener, in front of a computer screen. The seventy-two trials (thirty-six critical trials and thirty-six fillers) were presented on the screen one by one, in randomized order for each participant. The visual domains for the confederate listener were shown on a laptop placed in front of him. The computer screen and the laptop were positioned such that eye contact between the speaker and the listener was possible. E-prime 2.0 was used to run the experiment.

The instructions emphasized that speakers had to describe the target objects so that the listener could uniquely identify them, and mark them on a paper answering form. Given that the listener was presented with the same visual domains as the speaker, but with the objects in a different configuration, the instructions for the speaker also mentioned that it would not make sense to refer to objects with location information. Irrespective of the condition that speakers were assigned to, they could take as much time as necessary for every trial to describe the target. Once the listener had identified a target, he pressed the space bar to continue to the next trial. Before each trial, a fixation cross was depicted in the middle of the screen for 1 second. There were three practice trials, including one with Greebles, and the speakers' descriptions were recorded with a microphone. The confederate listener never asked clarification questions, so the data presented here can be regarded as one-shot reference.

In order to measure if speakers in the system-paced condition indeed experienced more pressure than those in the self-paced condition, all speakers filled out a short questionnaire after finishing their referential task. This questionnaire consisted of five questions on a 10-point scale, derived from the NASA task load index (Hart & Staveland, 1988). We asked

speakers to estimate: the task load (1); the pressure that they experienced (2); how well they had succeeded in describing the objects (3); how hard they worked to describe the targets (4); and how frustrated they were during the experiment (5). After reversing the scores for the second question and calculating the mean for each speaker, we found that the speakers in the system-paced condition (M = 4.62; SD = 1.54) indeed experienced their task as more demanding than the speakers in the self-paced condition (M = 3.37; SD = 1.24), F(1,34) = 7.27; p = .011).

Data coding

In total, our 36 speakers produced 1296 object descriptions. There were 25 missing cases due to technical issues with the audio recordings. The remaining 1271 descriptions were all coded for the occurrence of referential overspecification. A description was either overspecified (score = 1) or not (score = 0). Overspecified descriptions always contained color *and* size: one attribute which was necessary for identification of the target referent, and one redundant attribute. Descriptions that were coded as not overspecified were mostly minimally specified, and contained either color *or* size. References that were underspecified (16 in total) were coded with 0 as well. We did not consider whether speakers used a type attribute or not in coding the references, because the seven objects in the domains were always of the same type. Naturally, type attributes were often mentioned by our speakers to produce proper noun phrases.

Research design and statistical analysis

The experiment had a 3x2 design, with *Condition* (levels: **C**, **S**, **C/S**) as a within-participants factor and *Pace* (levels: self-paced, system-paced) as a between-participants factor. The dependent variable was the proportion of object descriptions that was overspecified.

To test the effect of Pace and Condition on the occurrence of referential overspecification, we performed a logit mixed model analysis (Jaeger, 2008). In our model, Pace and Condition were included as fixed factors, and items and participants as random factors. The fixed factors were centered to reduce collinearity. The model had a maximal random effect structure and included random intercepts and random slopes for all within-participant and within-item factors, in order to ensure optimal generalizability (Barr, Levy, Scheepers & Tily, 2013). As such, the model contained random intercepts for participants and items, and random slopes for Condition and Pace at both the participant level and the item level. The p-values were estimated via parametric bootstrapping over 100 iterations.

Results

Figure 2 plots the proportion of descriptions containing only color, only size, or color-and-size, as a function of the three Conditions (**C**, **S**, **C/S**) and the manipulation of Pace.

Overall, 36% of the referring expressions produced by our speakers contained both a color and a size attribute and were thus overspecified. Our model showed a main effect of Pace

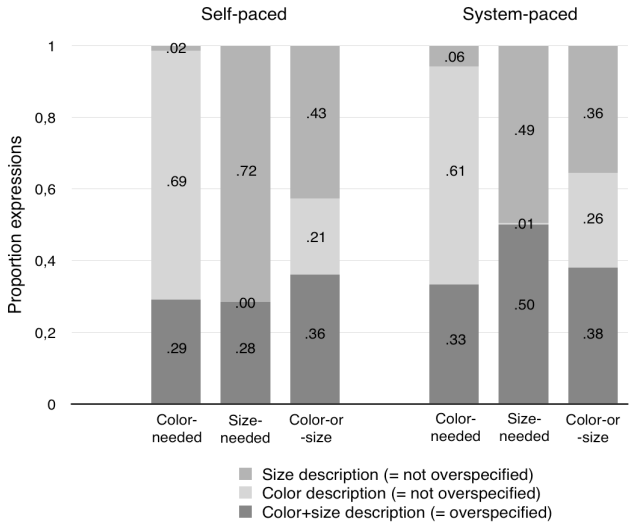


Figure 2: Proportion of color-only, size-only, and color-and-size descriptions as a function of the three Conditions and the Pace manipulation. Note that only the color-and-size descriptions (see the lower, darker bars) are overspecified.

on the occurrence of referential overspecification: redundant modifiers were more frequent in the system-paced condition (40.7%) than in the self-paced condition (31.3%), β = 0.72; SE = 0.57; $p < .05$. As can be seen in Figure 2, this effect of Pace was consistent across the three Conditions: the patterns for the **C**, **S**, and **C/S** conditions separately all show a higher proportion of overspecification for speakers who had limited time to inspect the scene. There was no main effect of Condition on the occurrence of overspecification (β = -0.01; SE = 0.36; n.s.).

Further inspection of the proportions in Figure 2 suggests that the effect of Pace is mediated by Condition. For instanc ple, in the **S** condition - where only size is needed to uniquely identify the target - the difference between speakers in the self-paced (28%) and the system-paced (50%) conditions is considerable in terms of overspecification, or at least bigger than in the **C** and **C/S** conditions. However, the interaction between Pace and Condition did not reach significance (β = 0.40; SE = 0.43; n.s.). This lack of interaction may be due to a substantial amount of speaker variation. We come back to this issue in the Discussion.

Discussion

The current paper investigated the effect of viewing time on attribute selection in the production of definite reference. In particular, the data revealed that viewing time influences the occurrence of referential overspecification: redundant modifiers were more frequent in the system-paced condition than in the self-paced condition. This effect did not interact with the type of visual domain, which we manipulated in terms of the attributes that were needed to uniquely identify the target referent (i.e., color, size, or color/size).

The main effect of *Pace* is in line with our expectations, and we regard it as converging evidence for the notion that speakers use quick heuristics when selecting the content of their referring expressions (e.g., Van Deemter et al., 2012; Dale & Viethen, 2009). In fact, speakers seem to particularly do so when their time to inspect the domain is limited, as was the case in our system-paced condition. The higher proportion of overspecified descriptions that we observed there suggests a heuristic, preference-based approach for speakers under pressure, where they select attributes merely based on inherent salience rather than discriminatory power. After all, these speakers may simply have lacked time and thus cognitive capacity (Horton & Keysar, 1996) to perform a deliberate object-by-object scan of the domain. Instead, they might have selected the content of their referring expressions in an incremental way (Pechmann, 1989). As we have seen in the Introduction section, this incremental process may eventually cause speakers to overspecify.

The lower proportion of overspecified object descriptions produced in the self-paced condition implies that a heuristic strategy is less dominant when speakers have unlimited time to inspect the visual scene. In this condition, there seems to have been sufficient time available for speakers to put effort in selecting the attribute that was most efficient in ruling out the distractor objects that were present. This way, speakers were – at least to a certain extent – able to avoid overspecification. However, it is important to note that if speakers in the self-paced condition had based the selection of attributes on discriminatory power alone, they may not have overspecified their descriptions at all, since there was in all domains one attribute that excluded all distractors at once. Hence, we take the finding that speakers did not avoid being redundant here as yet another argument for a model of human attribute selection where heuristic and more deliberate processes take place in parallel. As referred to in the Introduction section, such a model is similar to various computational models of attribute selection for definite reference (Gatt et al., 2011a; Mitchell et al., 2013).

The data did not show a significant effect of *Condition* on the occurrence of overspecification, in contrast to the related experiment by Gatt et al., 2011b, which manipulated similar conditions. This lack of main effect can be explained by the fact that in our study, the use of basic color terms was disabled, in an attempt to balance the trade-off between speaker preferences for color and size attributes. Normally, speakers have a strong preference to use color (e.g., Pechmann, 1989; Koolen et al., 2013), which could result in a high proportion of color-only descriptions in the **C** and **C/S** conditions, and a ceiling effect for overspecification in the **S** condition, even for self-paced speakers. In other words: speakers might then simply select color all the time, just because it is highly salient. However, with our subtle manipulation of color differences, we managed to discourage this strategy.

Close inspection of Figure 2 suggests the existence of an interaction between *Pace* and *Condition*: at least numerically, the effect of pressure was far most convincing in the **S** condition, where size was needed to distinguish the target. In this condition, the proportion of overspecifications almost doubled for speakers under pressure, while there was only a slight increase in the **C** and **C/S** conditions. However, the interaction between Pace and Condition was not significant. As mentioned earlier, we expect that this lack of interaction was due to a substantial amount of speaker variation, especially in the self-paced condition. The pattern that emerges there is that for almost half of the speakers, we find a proportion of 0 - 10% overspecified references, while the proportions for the other speakers range from 10 - 100%. In the system-paced condition, these proportions are more centered around the mean. This difference in consistency between the two conditions could explain why the interaction between pressure and condition was not significant.

Although the interaction effect was not significant, it remains of course interesting to speculate about the nature of the large numerical increase in the number of overspecified descriptions in the **S** condition. We reason that the inherent preference for color over size pays off here anyway, in spite of the subtle color differences in our domains. In cases with unlimited viewing time, color differences are generally detected faster than size differences (Belke & Meyer, 2002), because color is an absolute attribute that does not require comparison to other objects to be perceived, unlike the relative attribute size. In our experiment, this natural difference between size and color may explain the large increase in the proportion of overspecified descriptions in the **S** condition: speakers under pressure could simply have lacked sufficient time to compare the objects and sizes. Selecting color could provide a solution here, since it reduces the distractor set to three rather than six objects. This subset may be sufficiently small for speakers to detect if also size is needed to uniquely identify the target, perhaps even after the domain has disappeared, based on memory. In order to test this idea, it would be interesting to replicate our experiment in an eye-tracking paradigm, to test if speakers indeed rely on a subset of distractors before (and after) the domain has disappeared.

Related to the above, another next step could be to look at the way in which viewing time affects reference production at the surface, word level. In Dutch, which was the language of the experiment, color and size are usually realized before the head noun. However, if speakers indeed mentioned color to reduce the number of distractors, and 'decided' about the need for size later, one may expect that many size modifiers in the system-paced condition were produced *after* the head noun. Hence, there could also be more speech repairs in this case, with speakers producing descriptions such as "the light green light bulb, the large one". Third, there may be effects of pressure on speech onset times, which might be longer in the self-paced rather than the system-paced condition. These kinds of analyses can all be conducted with the current data; we are planning to do so in the near future. Hence, it would also be interesting to replicate our experiment in a language where post-nominal modification is the default, which is for example the case in Spanish or Maltese.

Other directions for future research include for example to run an experiment in which one of the attributes rules out no

distractors at all. In all conditions of the current experiment, both color and size ruled out at least one distractor object. In the system-paced condition, speakers might have included a redundant attribute because – during their quick scan of the scene – they realized that it excluded at least one distractor, although they were not sure how many distractors. If such a strategy is indeed applied, limited viewing time might not give rise to overspecification when certain attributes rule out no distractors at all, since during the quick scan, the speaker will not identify any distractor that these attributes would rule out. Second, one could conduct a follow-up experiment in which it is crystal clear for speakers that minimality (i.e., producing the shortest possible distinguishing description) is the aim, while giving them a very short inspection time. It is then the question to what extent their referential behavior resembles the behavior in the normal situation, where minimality is not particularly stressed. Finally, it would be interesting to explore the interplay between limited viewing time and overspecification from a more listener-oriented perspective, for example with a colorblind listener.

Conclusion

This paper explored the impact of viewing time on attribute selection in definite reference, and on referential overspecification in particular. We found more redundant attributes in the system-paced condition than in the self-paced condition, and conjectured that speakers rely heavily on fast heuristics when they have limited time to inspect the visual domain. In the case of unlimited inspection time, they seem to rely on a combination of heuristic and more deliberate strategies.

References

Barr, D., Levy, R., Scheepers, C. & Tily, H. (2013). Random effects structure for confirmatory hypothesis testing: keep it maximal. *Journal of Memory and Language, 68*, 255– 278.

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.

Brennan, S. & Clark, H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory & Cognition, 22*, 1482-1493.

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

Dale, R. & Viethen, J. (2009). Referring expression generation through attribute-based heuristics. *Proceedings of the 12th European workshop on natural language generation (ENLG)*. Athens, Greece, 58–65.

Gatt, A., Goudbeek, M. & Krahmer, E. (2011a). Attribute preference and priming in reference production: Experimental evidence and computational modeling. *Proceedings of CogSci 2011*. Boston, Massachusetts, 2627-2632.

Gatt, A., Van Gompel, R., Krahmer, E. & Van Deemter (2011b). Non-deterministic attribute selection in reference production. In *Proceedings of PRE-CogSci 2011*. Boston, Massachusetts.

Gauthier, I. & Tarr, M. J. (1997). Becoming a "Greeble" expert: Exploring mechanisms for face recognition. *Vision Research*, *37*(12), 1673–1682.

Hart, S., & Staveland, L. (1988). Development of NASA-TLX: results of empirical and theoretical research. In: Hancock, P.A., Meshkati, P. (Eds.), *Human Mental Workload* (pp. 139-183). Elsevier, Amsterdam.

Horton, W. & Keysar, B. When do speakers into account common ground? *Cognition, 59*, 91-117.

Jaeger, T. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59*, 434-446.

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

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

Mitchell, M., Van Deemter, K. & Reiter, E. (2013). Generating expressions that refer to visible objects. In *Proceedings of the North American chapter of the Association for Computational Linguistics: Human Language Technologies*. Atlanta, Association for Computational Linguistics.

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

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.

Tarenskeen, S., Broersma, M. & Geurts, B. (2015). Overspecification of color, pattern and size: salience, absoluteness, and consistency. *Frontiers in Psychology, 6*: 1703.

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

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

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 Gompel, R., Gatt, A., Krahmer, E. & Van Deemter, K. (2014). Overspecification in reference: modelling size contrast effects. *Poster Presented at AMLaP 2014* (Edinburgh, UK).

Viethen, J., Goudbeek, M. & Krahmer, E. (2012). The impact of colour difference and colour codability on reference production. *Proceedings of CogSci 2012*. Sapporo, Japan, 1084-1098.