Effects of scene variation on referential overspecification

Ruud Koolen (R.M.F.Koolen@uvt.nl)
Martijn Goudbeek (M.B.Goudbeek@uvt.nl)
Emiel Krahmer (E.J.Krahmer@uvt.nl)
Tilburg centre for Cognition and Communication (TiCC), Tilburg University
P.O. Box 90153, NL-5000 LE Tilburg, The Netherlands

Abstract

This study presents the results of two experiments conducted to investigate how the amount of variation between target and distractor objects in a visual scene influences referential overspecification. We hypothesized that as this variation gets higher, speakers tend to include more redundant information in their target descriptions. The results showed that this was indeed the case. We suggest that scene variation causes speakers to make use of quick heuristics when selecting the content of their referring expressions, and discuss the implications of these findings for computational models that automatically generate referring expressions.

Keywords: Referential overspecification; scene variation; computational models.

Introduction

In everyday language use, speakers often refer to objects in the world (*target* objects) in such a way that an addressee is able to uniquely identify them from other objects (*distractor* objects). A common way to do this is by producing definite descriptions of the target, such as '*the green chair*' or '*the brown desk*'. In order to make communication successful, speakers constantly need to decide on the semantic content of these referring expressions. Bach (1994) argued that referring expressions are contrastive by means of their *distinguishing attributes*; these are attributes that can be ascribed to the target, but not to the distractors. Still, the question remains which and how many target attributes should be provided to make the target easily identifiable.

It is often assumed that speakers tend to obey the Maxim of Quantity by Grice (1975), stating that speakers should be 'only as informative as required'. This would result in minimally distinguishing target descriptions, containing just enough information (i.e., target attributes) for successful target identification, but not more information. However, several psycholinguistic studies have shown that speakers do not always follow Grice's Maxim of Quantity, and that they overspecify their referring expressions by including *redundant* attributes of the target. In other words, referring expressions often contain information that is not needed for unique target identification (Pechmann, 1989; Maes, Arts & Noordman, 2004; Engelhardt, Bailey & Ferreira, 2006). Why would speakers do this so often?

It is generally assumed that referential overspecification is guided by both speaker- and addressee-oriented processes (Arnold, 2008). Several studies have revealed that listeners find it easier to identify a target when they are provided with an overspecified reference rather than a minimally specified one (e.g., Nadig & Sedivy, 2002; Paraboni, van Deemter & Masthoff, 2006), and that overspecified expressions often lead to shorter identification times (Arts et al., 2011). While this addressee-oriented approach is considered to be the traditional cognitive view on overspecification, the focus in this paper lies on the *speaker-oriented* processes that cause speakers to overspecify. Experimental evidence for the occurrence of these processes comes from Belke and Meyer (2002), who show that speakers tend to include absolute and perceptually salient attributes (such as 'colour') in their referring expressions, even if these attributes do not have contrastive value. This result is in line with earlier work by Eikmeyer and Ahlsèn (1996), Pechmann (1989), and Schrieffers and Pechmann (1988), who also found that speakers tend to include perceptually salient attributes in their target descriptions, even if these attributes do not directly serve the target identification goal.

The above psycholinguistic considerations concerning the occurrence of referential overspecification have important implications for researchers in the field of Natural Language Generation (NLG), who build systems that automatically generate natural language text or speech from non-linguistic information (e.g., from a database; Reiter & Dale, 2000). NLG systems typically require Referring Expression Generation (REG) algorithms that automatically generate distinguishing descriptions of objects (Mellish et al., 2006). Various REG algorithms have been proposed, and many have taken the Maxim of Quantity as a starting point (Dale & Reiter, 1995). For example, the Full Brevity Algorithm (Dale, 1989) is based on a strict interpretation of the Maxim and seeks to find the shortest possible target description (in terms of the number of attributes included). The Incremental Algorithm (Dale & Reiter, 1995) proposes a more relaxed interpretation of the Maxim of Quantity, since it attempts to account for the occurrence of referential overspecification by using a predetermined *preference order* for all possible target attributes in a particular domain. In practice, this means that the target's 'type' is added first. In case this leads to a distinguishing description, the system terminates the expression without including any other target attributes. If 'type' does not rule out all distractors, preferred target attributes such as 'colour' are added to the description (but only if they have some contrastive value). If that still does not suffice, less preferred attributes such as 'size' are added. However, since the algorithm does not backtrack for redundancy, it does not remove preferred attributes that turn out to be redundant in the end (because there turns out to be another - less preferred - attribute that excludes all

distractors at once). In this way, overspecified expressions can be generated.

The question is to what extent the current REG algorithms are psychologically realistic. Van Deemter et al. (accepted) argue that REG algorithms have several properties that are problematic in this respect. First, REG algorithms are typically deterministic, that is, they always generate the same referring expression in a particular context. Obviously, this is not in line with what humans do. Second, although we have seen that some of the current REG algorithms are somehow able to deal with referential overspecification, they have not found a systematic way to do this. Arguably, human speakers seem not to have such problems, suggesting that they rely on different (and more clever) attribute selection mechanisms that the current REG algorithms do not yet incorporate. More concretely, it is argued by many researchers that people may rely on *quick heuristics* when making decisions (Tversky & Kahnemann, 1982). Van Deemter et al. (accepted) suggest that similar processes might also play a role when speakers produce referring expressions: instead of searching for attributes with high distinguishing value, they may base their attribute selection on other criteria (e.g., attribute preference).

In this paper, we aim to investigate in more detail the differences between human heuristics and the mechanisms that REG algorithms base their content selection on. We assume that these differences become larger when the referring task gets more difficult, in particular when the *scene variation* gets higher: that is, when the objects in a scene differ along a higher number of attributes. Our central hypothesis is that a high scene variation causes speakers to overspecify their references more frequently, as compared to when this variation is low. More specifically, we expect that in situations where the scene variation is high, heuristics cause human speakers to include preferred - but redundant - attributes in their descriptions, and that this causes the expressions to be more frequently overspecified. When confronted with a simple scene, speakers might be more likely to be able to quickly determine which attributes distinguish a target from its distractors. However, when a scene becomes more varied (and hence more complex) speakers might be more likely to rely on a heuristic, which causes them to select attributes from the target without making sure that these are strictly needed to distinguish the target. Based on prior research discussed earlier, it seems plausible that speakers for this will prefer absolute (such as colour) over relative attributes (such as size). We therefore expect our participants to use colour more frequently (even when it is redundant) when a visual scene displays a high variation than when it does not.

In contrast, the current REG algorithms act differently in the exact same communicative situations as compared to humans, by generating minimally distinguishing referring expressions instead of overspecified ones, irrespective of the variation in the scene. For example, in situations where 'type' would be sufficient to distinguish a particular target, for example the Full Brevity Algorithm and the Incremental Algorithm would never include redundant attributes in their descriptions. Thus, if humans would indeed overspecify their references more when the scene variation gets higher, improvement of the current REG algorithms is needed to make their output more psychologically realistic.

In order to investigate the effect of scene variation on the amount of referential overspecification, we performed two experiments in which participants were presented with picture grids consisting of eight pictures (one target and seven distractors), asking them to produce distinguishing descriptions of the target objects. These two experiments consisted of two conditions: one in which the variation in the scene was kept low, and one in which the variation was high. The amount of variation in the scenes varied between experiments: in Experiment 1, targets could be distinguished in terms of their type only, while in Experiment 2 additional attributes were required. In neither of the experiments, 'colour' was needed to distinguish the target. We will study whether a higher scene variation indeed causes speakers to include more redundant target attributes in their referring expressions. Finally, we will contrast our findings with the state-of-the-art REG algorithms.

Experiment 1

Method

Participants Participants were undergraduate students who participated in pairs. Twenty-one students (10 male, 11 female, mean age = 21 years and 7 months) acted as speakers in this experiment. Another twenty-one students acted as addressees. All participants were native speakers of Dutch and participated for course credits.

Materials The stimulus material consisted of artificially constructed pictures of furniture items[1], which have been extensively used before in the field of REG generation (i.e. Gatt et al., 2007). The furniture items varied in terms of four attributes and their corresponding values. All possible attribute-value pairs are listed in table 1.

Table 1: Attributes and possible values of the furniture items.

Attributes	Possible values
type	chair, sofa, fan, television, desk
colour	red, blue, green, brown, grey
orientation	front, back, left, right
size	large, small

The critical trials all contained eight furniture items: one target object and seven distractor objects. The basic idea of the experiment was that two participants took part in a language production task, where one participant (the

[1] These objects were taken from the Object Databank, developed and freely distributed by Michael Tarr at Brown University. URL: http://www.tarrlab.org/

speaker) provided descriptions of the target objects and the other one (the *addressee*) used these descriptions to identify the corresponding target objects by distinguishing them from the distractor objects. For the speakers, the target referents were clearly marked by black borders so that they could easily distinguish them from the distractor objects. The furniture items were positioned on a computer screen in a 2 (row) by 4 (column) picture grid.

Experiment 1 had two conditions. The critical trials in the *low variation condition* were constructed in such a way that there was limited variation between the target and the distractor objects: the furniture items differed only in terms of the attribute 'type'. This means that the grid contained different types of furniture items that all had the same colour, orientation and size. In the *high variation condition*, however, the target and the distractor objects differed in terms of all four possible attributes: 'type', 'colour', 'orientation' and 'size'. Mentioning 'type' was sufficient to successfully distinguish the target in all critical trials in the two conditions, which implies that including preferred attributes such as 'colour' was never needed to distinguish the target. Figure 1 depicts examples of critical trials in the two respective conditions.

The trials were built in such a way that an algorithm like the Incremental Algorithm would never include 'colour' in its descriptions: since mentioning 'type' was sufficient for distinguishing the target in both of the two conditions, the algorithm would not include any further preferred (but redundant) attributes.

Figure 1: Examples of critical trials in Experiment 1: for the low variation condition (upper picture) and for the high variation condition (lower picture). Manipulations of colour may not be visible in a black and white print of this paper.

There were ten critical trials in each of the two conditions, giving rise to twenty critical trials. Together with forty fillers, this made a block of sixty trials in a fixed random order, which was counterbalanced for order across the experiment. The fillers consisted of four pictures of Greebles (Gauthier & Tarr, 1997): one clearly marked target referent and three distractor objects, all positioned in a 2 by 2 picture grid. Because initially designed so as to share characteristics with human faces, Greebles are complex and difficult objects to refer to, which made them useful fillers in our experiment. The Greebles could not be distinguished in terms of their colour because they were all in the same colour every time (so speakers were not primed with the attribute colour when describing the fillers).

Procedure The experiment was performed in an experimental laboratory. After the two participants had arrived in the room, it was randomly decided who was going to act as the speaker and who as the addressee, whereafter they were seated opposite to each other. The speaker was presented with the sixty trials on a computer screen, and was asked to describe the target referents in such a way that the addressee would be able to uniquely identify them. The instructions emphasized that it would not make sense (and that it was not allowed) to include location information in the descriptions, since the addressee was presented with the pictures in a different order. The speaker could take as much time as needed to describe the target, and his or her target descriptions were recorded with a voice recorder. The addressee was presented with the same sixty trials as the speaker in a paper booklet, and was asked to mark the picture that he or she thought the speaker was describing on an answering form. The instructions emphasized that the addressee was – to a limited extent – allowed to ask for clarification: it was allowed to ask the speaker to give more information or to repeat information that had already been given, but not to ask for specific information (i.e., specific attributes). Once the addressee had identified a target, this was communicated to the speaker, who then went on describing the next one. After completion of the experiment, none of the participants indicated that they had been aware of the actual goal of the study.

Design and statistical analysis Experiment 1 had a within participants design with Scene variation (levels: low, high) as the independent variable, and the average number of referring expressions containing a colour attribute (as explained below) as the dependent variable. Our statistical procedure consisted of two repeated measures ANOVAs: one on the participants means with the participants as the random variable ($F1$), and one on the item means with the items as the random variable ($F2$).

We chose the proportional use of the attribute 'colour' as the dependent variable indicating overspecification in referring expressions. As described above, we made sure that speakers never needed to include colour in their descriptions in order to produce a distinguishing description of the target. Thus, if speakers mentioned colour anyway, this caused the expression to be overspecified.

Results

Figure 2 depicts the proportion of expressions that contained a colour attribute as a function of the condition in which the descriptions were uttered.

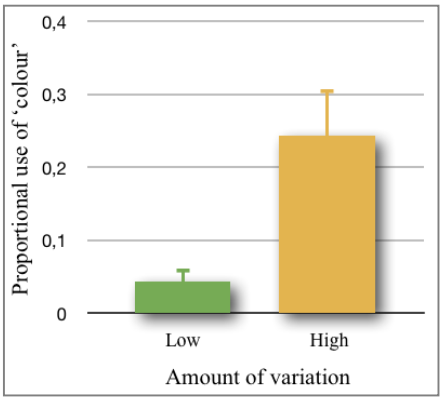

Figure 2: The proportion of referring expressions (plus standard deviations) containing a 'colour' attribute as a function of the variation in the visual scene.

As hypothesized, the scene variation affected the proportional use of the redundant attribute 'colour' ($F1_{(1,20)} = 12.537$, $p = .002$; $F2_{(1,18)} = 23.416$, $p < .001$). More specifically, speakers were more likely to include 'colour' when there was high variation in the picture grid ($M = .24$, $SD = .07$) compared to when this variation was low ($M = .04$, $SD = .02$).

This first experiment confirmed our hypothesis about the role of scene variation on speakers' tendencies to include redundant attributes in their target descriptions. In the next experiment, we will see whether the same applies when the difference between the low and high variation conditions gets more subtle.

Experiment 2

Method

Participants Participants were again undergraduate students who participated in pairs. This time, there were twenty-two students who acted as speakers (10 male, 12 female, mean age = 22 years and 4 months). None of these speakers acted as a speaker in Experiment 1. Another twenty-two students acted as addressees in this experiment. Most of these had been speakers in Experiment 1, in a few cases the addressee was a confederate. The participants were all native speakers of Dutch and participated for course credits.

Materials Again, there were twenty critical trials in two conditions, and these trials all contained one clearly marked target referent and seven distractor objects. Like in the first experiment, we included forty fillers consisting of four pictures of Greebles (Gauthier & Tarr, 1997).

Again, there was maximum variation between the target and the distractor objects in the *high variation condition* (thus, the objects again differed in terms of the attributes 'type', 'colour', 'orientation' and 'size'). However, unlike in Experiment 1 (where the objects only had different types), the pictures in the *low variation condition* now varied in terms of three attributes: again 'type', but also 'orientation' and 'size'. This caused the difference between the trials in the two conditions to be more subtle as compared to in Experiment 1.

Figure 3 depicts examples of trials in the two conditions of experiment 2. In all critical trials, mentioning 'type' plus one other attribute ('orientation' or 'size') was sufficient to produce a distinguishing description of the target. Again, mentioning 'colour' was never needed to distinguish the target. As in Experiment 1, the trials were built in such a way that algorithms like the Incremental Algorithm would never include 'colour' in their target descriptions. In the low variation condition in figure 3, the algorithm would not select 'colour' because all pictures have the same colour. In the high variation condition in figure 3, the algorithm will

Figure 3: Examples of critical trials in Experiment 2: for the low variation condition (upper picture) and for the high variation condition (lower picture). Manipulations of colour may not be visible in a black and white print of this paper.

first select 'type'. Since both remaining objects are then brown chairs, the algorithm will then select 'size' instead of the preferred attribute 'colour'.

Procedure, design and statistical analysis As above.

Results

Figure 4 depicts the proportion of expressions that contained a colour attribute as a function of the condition in which the descriptions were uttered.

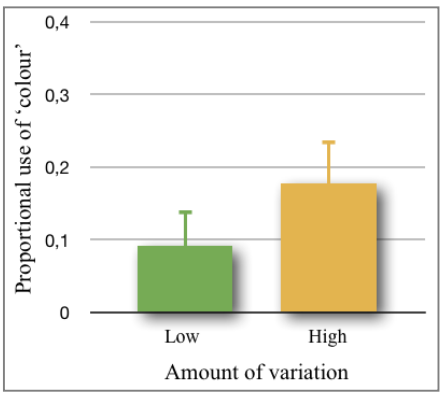

Figure 4: The proportion of referring expressions (plus standard deviations) containing a 'colour' attribute as a function of the variation in the visual scene.

The general picture of the results of this experiment is comparable to that of the results of Experiment 1. We again found that the amount of variation between the target referent and the distractor objects affected the number of times that speakers included the redundant attribute 'colour' in their referring expressions ($F1_{(1,21)}$ = 7.092, p = .015; $F2_{(1,18)}$ = 10.515, p = .005). More specifically, the results showed that speakers were more likely to include 'colour' when the variation in the picture grid was high (M = *.18, SD* = *.06*) compared to when it was low (M = *.09, SD* = *.05*).

The results of Experiment 2, as those of Experiment 1, confirmed our central hypothesis, and indicate that speakers include more redundant attributes in their references when the variation in a visual scene is high.

Discussion

The results of the two experiments presented in this paper show that when the objects in a visual scene vary along a relatively high number of dimensions, speakers are more likely to mention such variant attributes when describing a target object (even if this would result in overspecification). In particular, our results show that this is true for 'colour': the participants more often mentioned colour when the target object occurred in a high variation scene, as compared to when it occurred in a low variation scene. Since all trials were constructed in such a way that mentioning 'colour' was never needed to distinguish the target from the distractors, these results suggest that a high variation in a visual scene leads to more overspecification in speaker's descriptions.

As we have seen in the introduction of this paper, several papers in psycholinguistics (e.g., Belke & Meyer, 2002; Pechmann, 1989) have shown that speakers tend to include absolute and salient attributes (such as 'colour') in their descriptions, even if these do not have constrastive value. Often, such redundant attributes cause expressions to be overspecified. However, as far as we know, none of the previous studies has studied whether overspecification is influenced by the amount of variation in the visual scene. Following Tversky and Kahneman (1982) and van Deemter et al. (accepted), one can argue that speakers are guided by *heuristics* when selecting the attributes that they want to include in their references. For example, one heuristic could be that speakers tend to mention attributes that vary along the objects in a scene. Our results suggest that this heuristic could count for salient attributes such as 'colour'.

The current state-of-the-art REG algorithms make use of other mechanisms than heuristics in order to select the content of their generated output. For example, the Incremental Algorithm proposed by Dale and Reiter (1995) sometimes includes redundant target attributes in referring expressions, but the mechanism that this algorithm rests upon in doing this (i.e. using a preference order) still differs from what humans do. The results of our two experiments provide evidence for this. In this respect, it needs to be emphasized that our experimental trials were constructed in such a way that the Incremental Algorithm would never include 'colour' in its generated descriptions. Our finding that speakers in both experiments often included 'colour' in their descriptions (contrary to the predictions of the Incremental Algorithm) underlines the differences between humans and algorithms in terms of the way in which they select the content of their expressions. This suggests that in order to generate *psychologically realistic* descriptions, the Incremental Algorithm needs to include preferred attributes such as 'colour' in its descriptions more often, even if they do not rule out any of the distractor objects.

The difference between making use of heuristics and the mechanisms that algorithms base their content selection on becomes even larger if one takes the relationship between overspecification and scene variation into account. As the results of our two experiments suggest, speakers are more likely to apply a heuristic when they are presented with a picture grid in which the variation between the target and the distractor objects is relatively high. For the Incremental Algorithm, this implies that it should be made sensitive to the variation in the scene in which the target occurs.

Our findings provide empirical evidence for a suggestion raised in a paper by Koolen, Gatt, Goudbeek and Krahmer (submitted), being that speakers include more (redundant) information in their referring expressions when the range of attributes that is available for a speaker to describe the target is high. Koolen et al. report the results of an experiment in which speakers were asked to produce referring expressions in two domains: furniture and people. The pictures in the

furniture domain were similar to the ones used in this study, and varied in terms of only four attributes ('type', 'colour', 'orientation' and 'size'). The pictures in the people domain, however, varied in terms of at least ten attributes (including 'age', 'hair colour', etc). The results revealed that references to people were indeed more frequently overspecified than references to furniture items. However, these results had one important restriction, namely that two different domains were compared. As a result, it could have been the case that the number of attributes that were available to describe a target object was not the only factor causing speakers to overspecify. In this paper, we have solved this restriction by comparing visual scenes within a single domain.

In future research, we aim to expand the current study by focusing on a characteristic of the above described people domain (as used by Koolen et al., submitted), namely that all pictures in this domain are of the same type (<type = person>). It can be argued that this specific characteristic may cause speakers to mention more redundant attributes, because the pictures look more perceptually similar (which could make the referring task more difficult). Therefore, we aim to compare the level of overspecification of target descriptions uttered in a visual scene consisting of furniture items of one type (e.g., eight desks) with target descriptions uttered in a scene consisting of objects of multiple types.

Conclusion

Speakers are more likely to include salient attributes such as 'colour' in their target descriptions when the variation in the visual scene they are presented with is high as compared to when it is low. Often, mentioning such attributes leads to overspecification, which is problematic for computational models that aim to generate psychologically realistic target descriptions.

Acknowledgments

The research reported in this paper forms part of the NWO VICI project 'Bridging the gap between psycholinguistics and computational linguistics: the case of referring expressions' (Grant 277-70-007). We thank Tessa Dwyer and Joost Driessen for assistance in collecting the data.

References

Arnold, J. E. (2008). Reference production: production-internal and addressee-oriented processes. *Language and Cognitive Processes, 23 (4)*, 495-527.

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

Bach, K. (1994). *Thought and reference*. Oxford University Press, Oxford.

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

Dale, R. (1989). Cooking up referring expressions. *Proceedings of the 27th annual meeting of the association for Computational Linguistics* (pp. 68-75). University of British Columbia, Vancouver, BC, Canada.

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

van Deemter, K., Gatt, A., van Gompel, R. & Krahmer, E. (accepted). Computational linguistics of reference production. Topics in Cognitive Science.

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

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

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

Eikmeyer, H. J. & Ahlsén, E. (1996). The cognitive process of referring to an object: A comparative study of German and Swedish. *Proceedings of the 16th Scandinavian Conference on Linguistics*. Turku, Finland.

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

Koolen, R., Gatt, A., Goudbeek, M., & Krahmer, E. Overspecification in referring expressions: causal factors and language differences. Submitted.

Maes, A., Arts, A. & Noordman, L. (2004). Reference management in instructive discourse. *Discourse Processes, 37*, 117-144.

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

Nadig, A. S., & Sedivy, J. C. (2002). Evidence of perspective-taking constraints in children's on-line reference resolution. *Psychological Science, 13*, 329-336.

Paraboni, I., van Deemter, K., & Masthoff, J. (2007). Making referents easy to identify. *Computational Linguistics, 33*, 229-254.

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

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

Schriefers, H. & Pechmann, T. (1988). Incremental production of referential noun phrases by human speakers. In: M. Zock, & G. Sabah (Eds.), *Advances in Natural Language Generation, volume 1*. Pinter, London.

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