The Impact of Colour Difference and Colour Codability on Reference Production

Jette Viethen (h.a.e.viethen@uvt.nl)

Martijn Goudbeek (m.b.goudbeek@uvt.nl)

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

Abstract

It has often been observed that colour is a highly preferred attribute for use in distinguishing descriptions, that is, referring expressions with the purpose of identifying an object within a visual scene. However, most of these observations were based on visual displays containing only colours that were maximally different in hue and for which the language of experimentation possessed basic colour terms. The experiment described in this paper investigates the question whether people's preference for colour is reduced if the colour of the target referent is similar to that of the distractors. Because colours that look similar are often also harder to distinguish linguistically, we also examine the impact of the codability of the used colour values. The results of our experiment show that, while people are indeed less likely to use colour when the colours in the display are similar, this effect is entirely due to the difficulty in naming similar colours. When the colours of target and distractors are similar but can be named using different basic colour terms, no reduction in colour use was observed.

Keywords: reference production, language production, colour

Introduction

Referring expressions are an essential part of communication. Whenever people engage in any type of discourse they use referring expressions to encode the entities that they are talking or writing about. Sometimes it suffices to use a pronoun to let the addressee know what is meant, but often a distinguishing description, a noun phrase differentiating the target referent from all other visually available distractor objects, is necessary. The production of such distinguishing descriptions has been a central theme for researchers both in psycholinguistic and in computational research on reference production. One particular question of interest is which attributes should be chosen for realisation in a distinguishing description, the problem of semantic content selection.

One of the most often made observations in psycholinguistic research regarding the choice of attributes for distinguishing descriptions is that people seem to favour colour over almost all other attributes when describing a target referent with the aim of identification (cf. Pechmann, 1989; Belke & Meyer, 2002; Sedivy, 2003; Brown-Schmidt & Tanenhaus, 2006; Arts, Maes, Noordman, & Jansen, 2011). This includes frequent redundant use of colour; cases in which the referring expression would be equally as distinguishing if colour was not mentioned. In some cases, people even use colour when all objects in a scene are of the same colour (Koolen, Goudbeek, & Krahmer, 2012).

However, as far as we know, all of this research was based on stimulus material using prototypical primary colours with clearly defined basic colour terms. In this paper, we investigate the question of whether people's preference for using the colour attribute diminishes or remains the same when the colour values in a visual scene are more similar to each other, and when no different basic colour terms exist for them.

Various researchers have argued that colour is preferred over, for example, size, in reference production, because it expresses absolute rather than relative information. In particular, Pechmann (1989) found in an early eye-tracking study that people usually begin to verbalise a description before they have fully scanned the scene. He found that a third of the descriptions in his data that contained both size and colour did not follow standard word order by mentioning colour before size (e.g., *the blue small car*).[1] He also noted that the first-mentioned attribute in overspecified descriptions was almost always colour, which often was ultimately not useful for the task of distinguishing the target referent from the visual context. He argued that both these observations might be due to the fact that colour is more easily cognisable than the other distinguishing features in his experiment because it can be perceived without having to compare the target referent to the other objects in the scene.

Belke and Meyer (2002) found similar overspecification effects for colour and size as Pechmann. They additionally provided eye-tracking evidence from a same–different judgement task for an account which credits this effect to differences in the way absolute and relative attributes are processed at a perceptual level. Based on experiments using the Stroop paradigm, Naor-Raz, Tarr, and Kersten (2003) even argued that an object's colour is an intrinsic component of the visual representation retained in long-term memory.

Another prominent source of evidence for people's preference for colour comes from corpus studies on purpose-built collections of referring expressions. The furniture section of the TUNA Corpus is a collection of human-produced distinguishing descriptions for furniture items differing in type, colour, size and orientation. In this corpus, colour is used redundantly more than three times as often as the other at-

[1] The standard word order is in this case identical for English and Dutch, the language of Pechmann's experiment.

tributes (Gatt, 2007, p. 82). In their recent experiments on semantic alignment in referring expressions, Goudbeek and Krahmer (2012) examined whether people can be primed to use a dispreferred attribute over a preferred one. Because they re-used the visual stimulus objects from the TUNA Corpus, they made the much higher frequency of colour over that of orientation in that corpus an underlying assumption in their experimental design.

A further corpus analysis by Viethen and Dale (2011), based on a large set of referring expressions for simple 3D scenes, also found that people mentioned the colour of the target object in a large proportion of the cases in which it was not necessary for identification. For size, on the other hand, their analysis found that its use depended highly on how well it distinguished the target from the other distractors, especially those of the same type as the target, pointing to a much more 'utilitarian' attitude towards size than towards colour. This is in line with findings from eye-tracking experiments which have shown that size is rarely used in situations where it adds no discriminatory power to the referring expression at all , while the same is not true for colour (Sedivy, 2003; Brown-Schmidt & Tanenhaus, 2006).

In light of this evidence, it is uncontroversial that colour plays a special role in referential communication. Yet, it must be noted that all of these results are based on stimuli with objects coloured in a small number of very different hues (red, blue, green, yellow, grey), sometimes even only black and white. In other words, the colour differences between the objects presented to participants were as large as possible.

No research exists using stimulus objects in similar colours. An intuitively plausible prediction is that the use of colour decreases as the similarity between the colours in the scene increases. This prediction follows also from Deutsch and Herrmann's (1976) third postulate (p. 43). They show that in a situation with two identical objects that only differ in width and height, with a large difference in width and a small difference in height, people tend to use only the width attribute in a distinguishing description, and vice versa. Herrmann and Deutsch extrapolate from these findings that, in any situation in which more than one attribute can be used for identification, people will tend to use the one in which the objects differ most. If, on the other hand, the observed high rate of colour use in referential communication is indeed due to a smaller cognitive effort involved in mentioning it, as many other psycholinguists and computational linguists have argued, it should be unchanged in situations with colours that are not maximally different.

A confounding factor lies in the varying codability of different colour values. The more similar two colours are, the more likely it is that they fall within the range of the same basic colour term, such as *red*, *blue* or *yellow*, depending on the basic colour terms that exist in a given language.[2] In such a case, more complex colour terms, such as *dark red* or *light blue* have to be constructed. It is conceivable that a colour value that is harder to encode is less likely to be verbalised.

Regarding the effect of colour difference, two conflicting hypotheses can be formulated:

1. The colour of an object is perceived independently from the colours of surrounding objects and gets included in distinguishing descriptions reflexively rather than based on a consideration of its usefulness. This hypothesis is in line with most claims in the literature and predicts that the extent of the difference between the target's colour and that of the distractor objects has no impact on people's reference behaviour.

2. The high use of colour is based to some extent on an assessment of the difference in colour between the target item and the distractors. Following from (Herrmann & Deutsch, 1976), a lower use of colour should be expected when the colours are similar than in situations where the colours are as different from each other as possible.

For the effect of colour codability our hypothesis is:

3. The codability of a target's colour with respect to distractor colours effects the likelihood of it being used in a distinguishing description. Colours that can be named by a basic colour term are more likely to be included than those for which a complex term has to be used.

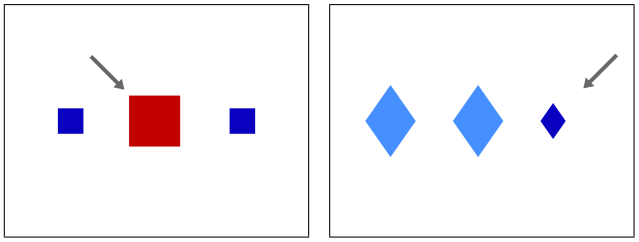
In the following, we describe an experiment designed to arbitrate between Hypotheses 1 and 2 and to test Hypothesis 3. Our results support the assumption of Hypothesis 3 that the use of colour is reduced when the codability of the colour value of an item is reduced, and advocate Hypothesis 1 over Hypothesis 2.

These results can inform ongoing research on developing computational models of reference production, as this work has begun to align its focus with that of psycholinguistic research. Researchers from the computational field are looking more and more for evidence about how humans solve the problem of content selection for reference production, in order to inform their models (cf. Dale & Reiter, 1995; Kelleher & Kruijff, 2006; Viethen & Dale, 2006; Deemter, Gatt, Sluis, & Power, 2012). One main reason for this move towards human-likeness as a criterion for task success of reference generation systems is the aim to create computational models that are in some sense cognitively plausible. The results of our experiment show that even computational models that are solely focussed on content selection for reference production need to pay more attention to the problem of lexical choice, as these two issues appear to be more closely intertwined than most existing models acknowledge.

Experiment

The experiment took the form of a reference production task, in which participants were shown displays of simple geometric objects on a computer screen. They were asked to describe

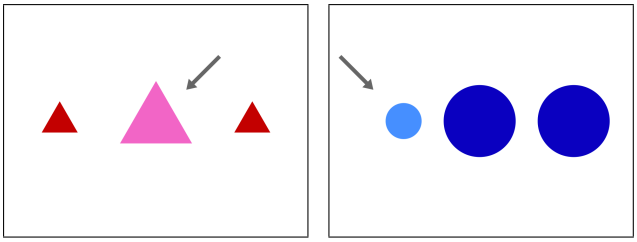
[2] Which hues are grouped under the same basic colour terms differs for different languages, as they carve up the colour spectrum in different ways and at different granularities (Kay, Berlin, Maffi, Merrifield, & Cook, 2010).



(a) A hidiff item: one large red and two small blue squares.

(b) A lodiff item: one small dark blue and two large light blue diamonds.

Figure 1: Example stimuli from the two colour-difference conditions.



(a) A hicode item: one large pink and two small red triangles.

(b) A locode item: one small light blue and two large dark blue circles.

Figure 2: Example stimuli from the two colour-codability conditions.

one of the objects in such a way that an imaginary partner would be able to identify it.

Method

Participants 63 undergraduate students of Tilburg University took part in the experiment in return for course credit. 48 were female and 15 male. Their age ranged from 19 to 26 years (M = 20 years and 10 months). They were all fluent speakers of Dutch, the language of the experiment.

Materials and Design Each participant was shown 32 critical items and 64 filler items. The critical trials consisted of simple scenes containing three two-dimensional geometrical figures: one intended referent and two distractor objects. In order to keep the design as simple as possible, the two distractor objects were identical. However, the target item differed in both colour and size from the two distractors, so that either of these two attributes was sufficient to fully distinguish it.

Our main manipulation concerned the difference in colour between the target and the distractors. In half of the trials this difference was large (hidiff condition), in the other half it was small (lodiff condition). Figure 1(a) shows a trial from the hidiff condition, and Figure 1(b) is an example from the lodiff condition.

As discussed above, the more similar two colours are, the less likely it is that they can be distinguished by basic colour terms. For example, the basic colour term *blue* is not sufficient to distinguish the target in Figure 1(b) from the distractors; instead, the complex colour term *dark blue* has to be used. This applies in Dutch in the same way as in English. To test the impact that the codability of different colour values might have on the content of referring expressions (see Hypothesis 3), we used a nested variable within the lodiff condition, by including two different hues: red and blue. For red hues, Dutch (just as English) possesses two different basic colour terms, even at a low difference, namely *rood* (red) and *roze* (pink). Thus, stimuli with red and pink objects, such as the one in Figure 2(a), form the hicode condition. For blue, the complex colour terms *donkerblauw* (dark blue) and *lichtblauw* (light blue) have to be used, resulting in a locode condition (an example stimulus is shown in Figure 2(b)). The lodiff items were equally divided between the hicode condition and the locode condition.

To determine the exact colour values to use we referred to the Hue Saturation Brightness (HSB) colour model. For the two dark colours we used the canonical values for blue (H = 245°) and red (H = 0°), 100% saturation, and a slightly lowered brightness (75%). For the lighter colours, we subtracted 35°from the original hue values, decreased the saturation and the increased the brightness. We finetuned the values for the lighter colours based on a pretest, to ensure that people would agree on calling them *roze* and *lichtblauw*. This resulted in the HSB values (215°, 70%, 100%) for light blue and (320°, 58%, 95%) for pink.

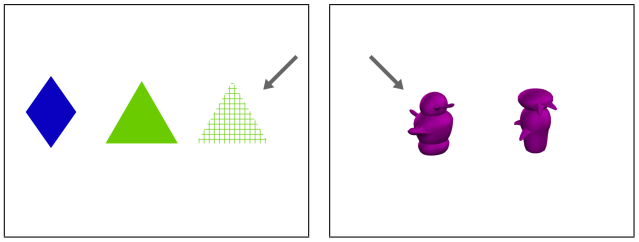
To ensure that there were the same number of target objects in each of the four colours (red, pink, dark blue and light blue), half the items in the hidiff condition used red and dark blue objects, and the other half pink and light blue ones. The position of the target was balanced across items. Furthermore, each condition contained a balanced number of trials using each of the four object types.

The type of the distractor objects was always the same as that of the target, so that type was never distinguishing. However, the size of the distractors was different from that of the target object, in order to give the participants an alternative option to using colour. It would not make sense to measure the rate of colour use, if colour was the only distinguishing feature in some or all trials.

We aimed to keep the size difference between target and distractors constant across all trials. To this end, we defined the size of an object by the length of its longest internal distance (the diameter for a circle, the diagonal for a square, an edge for a triangle, and the vertical line in a diamond), rather than, for example, its area. The longest internal distance of the large objects was set to twice that of the small objects.

Filler Items We included two types of fillers, which were carefully designed to mislead the participants regarding the exact aims of the experiment.

The 32 geometrical fillers were similar to the critical stimuli in that they showed three geometrical objects, but they used type and pattern as distinguishing attributes. Colour and size were never fully distinguishing in the fillers, in order to avoid priming the use of these two attributes. The target was



(a) A geometric filler. (b) A Greeble filler.

Figure 3: Two example filler items.

either striped or checkered so as not to prime the use of *solid* as a pattern which might then also show up in the critical trials where it was non-distinguishing. Half of the geometrical fillers were in black and white and in 9 of them the target was green, in order to distract from the small set of colours used in the critical trials. Again, the target referent's type and size was balanced across the whole set. Figure 3(a) shows an example of a geometric filler.

The 32 'Greeble' fillers each showed two novel 3D figures in purple.[3] We chose pairs such that the target object could always be distinguished from the distractor object by its main shape and the direction in which its protrusions were pointing. Because these objects are designed to be difficult to describe and look very different from the geometric items, we hoped they would prevent the participants from adopting a standard strategy for describing the geometric items. An example Greeble filler is shown in Figure 3(b). Debriefing revealed that the participants were not aware of the purpose of the experiment, and the majority of participants believed that the Greeble items were the critical stimuli of the experiment.

Procedure Two stimulus lists were created by producing one random ordering and then reversing it for the second list. Each critical stimulus was prepended with one geometrical and one Greeble filler item, which were chosen semi-randomly in a way such that the target was never in the same position in more than four items in a row. The item directly before each critical stimulus was always a Greeble filler to minimise any possibility of lexical or semantic priming from the geometrical filler responses to the critical responses.

The Dutch instructions told the participants that they would see a number of simple scenes on a computer screen. They were asked to verbally describe the object pointed at by an arrow to an imaginary partner without using position information. They had to complete the sentence *Klik nu op de/het* ... ('Now click on the ...') which was shown underneath each item. Their voice was recorded using a headset.

Before each item, a fixation cross was displayed for 1.5 seconds, then the stimulus item was shown for 4.5 seconds during which the participants had to give their response. We introduced this relatively short response time after finding in a pilot experiment that participants tended to exhaustively describe the whole scene.

[3]The Greebles are courtesy of Michael J. Tarr, Center for the Neural Basis of Cognition and Department of Psychology, Carnegie Mellon University, http://www.tarrlab.org/.

Table 1: Count and proportion of responses containing colour.

condition	count	mean	stdev
hidiff (N=1008)	747	.74	.35
lodiff (N=1008)	681	.68	.36
lodiff-hicode (N=504)	376	.75	.36
lodiff-locode (N=504)	305	.61	.40

Results

Coding of the Independent Variables As main dependent measure, we analyse the proportion of colour use in the different conditions. We consider a description to contain colour, if a colour value is mentioned that is true of the target, independently of the distinguishingness of this value. For example, we consider the description in Example (1) for the target in Figure 2(b), where all three objects are blue, as a use of colour. As a secondary measure we also look at the use of size, in order to get an insight into the question whether colour gets mostly used redundantly.

(1) de kleine blauwe cirkel

(the small blue circle) [for the stimulus in Figure 2(b)]

The responses were transcribed and coded for use of colour and size by a Dutch native-speaker.

Data Analysis Table 1 displays the mean proportion and standard deviation of colour use in the hidiff and lodiff conditions as well as the two nested conditions under lodiff (hicode and locode). It shows that people were more likely to use colour when the colours in the stimulus scene were very different than when they were similar. However, it also shows that the mean proportion of colour use in the lodiff-hicode condition was very similar to that in the hidiff condition.

We conducted a within-participants analysis of variance (ANOVA) to compare the three meaningful conditions (hidiff, lodiff-hicode, lodiff-locode), which showed the differences between these conditions to be highly statistically significant ($F(2,124) = 19.9, p < .001, \eta^2 = .24$). A test of planned within-participant Contrasts confirmed that the difference between the hidiff and the lodiff condition was significant ($F(1,62) = 18.7, p < .001, \eta^2 = .23$); participants used colour more in the hidiff condition than in the lodiff condition. The same is the case for the effect of codability (locode vs. hicode conditions) ($F(1,62) = 20.3, p < .001, \eta^2 = .25$), confirming that people used colour more when the colours could be distinguished by basic colour terms. However, there was no statistically significant difference between lodiff-hicode and the hidiff condition ($F(1,62) < 1$).

For size, the opposite picture emerges. Table 2 shows that people were less likely to use size when the colour difference was high than when the colours were similar, and that people used size more often in situations in which the name of the

Table 2: Count and proportion of responses containing size.

condition	count	mean	stdev
hidiff (N=1008)	584	.58	.36
lodiff (N=1008)	679	.67	.31
lodiff-hicode (N=504)	282	.56	.36
lodiff-locode (N=504)	397	.79	.30

colour was difficult to encode. Again, the difference between the hidiff and the lodiff-hicode conditions does not appear very big.

The statistical analysis with tests of planned Contrasts revealed the same pattern as for colour use: the overall difference between hidiff, lodiff-hicode, and lodiff-locode is significant with an even bigger effect size ($F(2,124) = 39.7, p < .001, \eta^2 = .39$); as are the differences between hicode and overall locode ($F(1,62) = 19.3, p < .001, \eta^2 = .24$) and between hicode and locode-lodiff ($F(1,62) = 52.3, p < .001, \eta^2 = .48$). This means that people used size less often in the hidiff and the hicode conditions than in the locode condition. Again, there was no statistically significant difference between hidiff and lodiff-hicode ($F(1,62) = 1.1$).

Discussion

The main observation from our results is that a smaller difference in colour alone does not result in a decrease in the use of colour in referring expressions. The apparent difference in colour use between the hidiff and lodiff conditions arises solely from the difficulty in coding the colour value in the locode condition. This lends support to Hypothesis 1, stating that people's preference for colour is independent from its value. It also confirms Hypothesis 3, which predicts that colours that are difficult to name because no distinguishing basic colour term is available, are less likely to be mentioned in a distinguishing description.

Interestingly, there were 99 distinguishing descriptions that contained a non-distinguishing colour value, such as in Example (1) above. All 99 of these cases occurred in the lodiff-locode condition. It is not surprising that no such cases occurred in the other conditions, because no basic colour terms exist that encompass both red and blue, red and pink, or pink and blue. However, the fact that almost a third of all colour terms used in the locode condition were non-distinguishing further supports the hypothesis that people often mention colour not for its discriminatory power but because it is easily available perceptually. By mentioning the basic, yet non-distinguishing, colour term *blauw* they can follow their preference for using colour but avoid the difficulty involved in retrieving and uttering a more complex colour term. This raises the question whether it is indeed the complexity of a colour term that stops people from using it or rather the fact that in our locode scenes the target's colour term (e.g. *lichtblauw* in Figure 2(b)) partly overlaps lexically with that applying to the distractors (*donkerblauw* in Figure 2(b)).

Furthermore, of the 37 descriptions in which a property was mentioned that was not true of the target object, only one used a wrong colour (*pink* instead of *red*, and in this case the participant corrected themselves). This further strengthens the argument that colour naming is an inherently easier task than naming the size made by a number of researchers including (Pechmann, 1989; Belke & Meyer, 2002; Naor-Raz et al., 2003; Kelleher & Kruijff, 2006).

The rate at which people used size was inversely proportional to the use of colour. Of course, size had to be used in descriptions not including colour in an identification task with these two attributes as the only distinguishing features. However, this does not necessarily mean that it has to be omitted in cases in which colour was mentioned. Instead, the rate of size use might have stayed constant, indicating a relatively high rate of overspecification in the hidiff and hicode conditions. Two possible explanations for the difference in the use of size between the different conditions are conceivable. First, it might be the case that the choice to use size is influenced directly by the experimental variables. This might be due to the fact that the speaker has to scan the scene in order to determine the relative size of the target object. While scanning a locode scene he might notice the usefulness of colour—which according to Pechmann's (1989) and Belke and Meyer's (2002) incrementality accounts might already have been uttered at this stage—and decide whether to use size based on this information alone, independently of whether colour is actually mentioned or not. Second, the use of size might be impacted by the use of colour. People might make their choices about which attributes to use sequentially, one attribute at a time and the decision about size succeeds the decision about colour. So, once a speaker has decided not to mention colour, size has to be included in order to fulfil the referential task of identification. Further experimentation would be required to arbitrate between these two accounts.

Consequences for Computational Modelling The main assumption regarding the use of colour remains unchallenged by our results: colour is highly preferred by human speakers and should therefore feature highly in the output of computational referring expression generation systems that are aimed at producing human-like output. However, our results re-emphasise the importance of an issue which seems to have lost traction in the decades since (Dale & Reiter, 1995): that of lexical choice. Dale and Reiter's original algorithm included a FindBestValue function, acknowledging the fact that different level values exist for many attributes and that not all values are equally adequate in a given situation. However, their algorithm makes its decision about which attributes to include based on the most distinguishing value for an attribute, meaning that a colour value expressed by a more complex term, such as *light blue*, is more likely to be included for the colour attribute than a basic one, such as *blue*. This is of course not advocated by our data.

Our findings speak loudly against the separation of semantic content selection and lexical choice present in most recent computational approaches to referring expression generation. Computational reference production models with a claim to

human-likeness need to take into account how difficult it will be to realise each attribute lexically already when they make the decision about the use of this attribute. The results presented here clearly show that even highly preferred attributes such as colour should get included less often in situations in which they are hard to code, or that in some cases a less specific value should get used.

A second point emerging from our data is that the deterministic nature of most existing computational reference production models is clearly not in line with human reference behaviour. While we can observe increases or decreases in the use of certain attributes depending on different experimental variables, there always remains a large amount of variation. Therefore, REG systems that are serious about modelling human behaviour must begin to use probabilistic mechanisms in order to be able to capture the non-deterministic choices people make when they refer. A notable first move in this direction was made by Gatt, van Gompel, Krahmer, and van Deemter (2011).

Conclusions

Previous research often took it for granted that colour is a highly preferred attribute in reference production, but so far a serious and systematic study of this has been lacking. Existing results were based on stimuli in maximally different primary colours; this paper is the first to investigate what happens if the stimulus colours are similar to each other. Our results suggest that the similarity between the colour of the target referent and that of any distractor objects indeed has little effect on the content people choose for a referring expression, supporting the view that colour gets chosen due to being perceivable with low cognitive effort.

However, we show that colours that can be encoded using a basic colour term, such as *blue*, are more likely to be mentioned than those for which a more complex term, such as *light blue*, has to be found in order to distinguish from, for example, dark blue distractors. Current computational models of reference production do not account for this result, as they usually separate the selection of semantic content and lexical choice into two distinct processes.

Acknowledgments

The research reported in this paper forms part of the VICI project "Bridging the Gap between Psycholinguistics and Computational Linguistics: the Case of Referring Expressions", funded by the Netherlands Organization for Scientific Research (NWO grant 277-70-007). We thank Elsa Jonkers for help with the transcription and annotation of the data.

References

Arts, A., Maes, A., Noordman, L., & Jansen, C. (2011). Overspecification in written instruction. *Linguistics*, *49*(3), 555–574.

Belke, E., & Meyer, A. S. (2002). Tracking the time course of multidimensional stimulus discrimination: Analysis of viewing patterns and processing time during same-different decisions. *European Journal of Cognitive Psychology*, *14*(2), 237–266.

Brown-Schmidt, S., & Tanenhaus, M. K. (2006). Watching the eyes when talking about size: An investigation of message formulation and utterance planning. *Journal of Memory and Language*, *54*, 592–609.

Dale, R., & Reiter, E. (1995). Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science*, *19*(2), 233–263.

Deemter, K. van, Gatt, A., Sluis, I. van der, & Power, R. (2012). Generation of referring expressions: Assessing the incremental algorithm. *Cognitive Science*, in press.

Gatt, A. (2007). *Generating Coherent Reference to Multiple Entities*. Unpublished doctoral dissertation, University of Aberdeen, UK.

Gatt, A., van Gompel, R., Krahmer, E., & van Deemter, K. (2011). Non-deterministic attribute selection in reference production. In *Proceedings of the Workshop on Production of Referring Expressions: Bridging the gap between empirical, computational and theoretical approaches to reference (PRE-CogSci 2011).* Boston MA, USA.

Goudbeek, M., & Krahmer, E. (2012). Alignment in interactive reference production: Content planning, modifier ordering and referential overspecification. *Topics in Cognitive Science*, *4*(2), 269–289.

Herrmann, T., & Deutsch, W. (1976). *Psychologie der Objektbenennung*. Bern: Verlag Hans Huber.

Kay, P., Berlin, B., Maffi, L., Merrifield, W. R., & Cook, R. (2010). *The World Color Survey*. Stanford CA, USA: CSLI Publications.

Kelleher, J., & Kruijff, G.-J. (2006). Incremental generation of spatial referring expressions in situated dialog. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics* (pp. 1041–1048). Sydney, Australia.

Koolen, R., Goudbeek, M., & Krahmer, E. (2012). The effect of scene variation on the redundant use of color in definite reference. *Cognitive Science*, to appear.

Naor-Raz, G., Tarr, M. J., & Kersten, D. (2003). Is color an intrinic property of object representation? *Perception*, *32*(6), 667–680.

Pechmann, T. (1989). Incremental speech production and referential overspecification. *Linguistics*, *27*(1), 89–110.

Sedivy, J. C. (2003). Pragmatic versus form-based accounts of referential contrast: Evidence for effects of informativity expectations. *Journal of Psycholinguistic Research*, *32*(1), 3–23.

Viethen, J., & Dale, R. (2006). Algorithms for generating referring expressions: Do they do what people do? In *Proceedings of the 4th International Conference on Natural Language Generation* (pp. 63–70). Sydney, Australia.

Viethen, J., & Dale, R. (2011). GRE3D7: A corpus of distinguishing descriptions for objects in visual scenes. In *Proceedings of the Workshop on Using Corpora in Natural Language Generation (NLG): Language Generation and Evaluation (UCNLG+Eval).* Edinburgh, UK.