Production of referring expressions: Preference trumps discrimination

Albert Gatt (albert.gatt@um.edu.mt)
Institute of Linguistics, University of Malta
Tilburg center for Cognition and Communication (TiCC), Tilburg University

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

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

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

Abstract

When referring to an object using a description, speakers need to select properties which jointly distinguish it from any potential distractors. Previous empirical and computational work addressing this content selection process has highlighted the role of both (i) the discriminatory power of properties of a referent, i.e. how many of the distractors in a domain each property excludes; (ii) how inherently salient or preferred a property is. To date, there has been no attempt to systematically investigate the trade-off between these two potentially competing motivations. This paper investigates experimentally the extent to which speakers take discriminatory power versus preference into account during content selection for reference production. Our results suggest that discriminatory power in fact plays a relatively unimportant role. We discuss the implications of this for computational models of reference production.
Keywords: Referring expressions, language production, psycholinguistics, computational modelling

Introduction

Referring expressions such as *the large bottle* are an essential feature of communication. Without the ability to refer, it would be difficult to ground our communicative efforts in the physical and mental world. The processes underlying reference production have been the object of intensive study by researchers in Computational Linguistics (see Krahmer & van Deemter, 2012, for an extensive review) and Experimental Psycholinguistics (e.g. Levelt, 1989; Arnold, 2008). Many researchers agree that the primary aim of a referring expression is to *identify* an object for an interlocutor, a position that is rooted in a long tradition of philosophical work on the subject (e.g. Searle, 1969).

Consider a situation in which a speaker needs to identify an object (the *target referent*), which has not been introduced earlier in the discourse and which is visually co-present for both speaker and listener. Here, the speaker needs to perform *content selection*, to determine which properties of the target referent to mention in a description. This process is non-trivial because objects have several properties to choose from; moreover, the goal of identification entails choosing a set of properties that jointly exclude all the *distractors* in the domain with which a listener might confuse the target. The speaker has to tread a fine line between efficiency on the one hand and sufficient detail on the other. Thus, it would seem desirable to avoid producing an *overspecified* description which contains more properties than required, or an *underspecified* one, which does not succeed in identifying the target. Both constraints would seem to follow to the extent that speakers observe Grice's Maxim of Quantity (Grice, 1975).

To take an example, the bottle with the black border in Figure 1 has three properties that are potentially distinguishing, namely, its colour, its size and the fact that it is marked with a black diamond (hereafter referred to as its *pattern*). On its own, none of these properties is sufficient to distinguish it from the distractors, the other bottles in the domain. Closer inspection reveals that this target minimally requires two properties (in fact, any two of these three) for successful identification. For example, *the large bottle with a diamond* would do the trick without overspecifying.

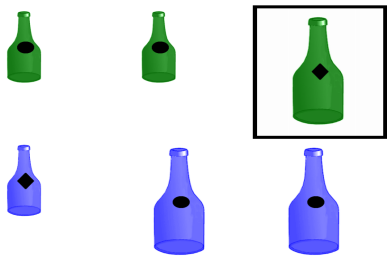

Figure 1: Example domain: the target's diamond pattern excludes 4 distractors, while its green colour excludes the 3 blue bottles on the bottom row

Models of content selection

What process would best model speakers' content selection procedure? It is widely accepted that, since speech production is incremental (cf. Levelt, 1989; Pechmann,

1989), properties would be selected one after the other. The main question is: on what basis is the choice made at each point? One possibility would be for the speaker to weigh each property in terms of its discriminatory power. For instance, looking at Figure 1, it is easy to see that starting with the target's pattern would eliminate four distractors, while either of the other two properties would eliminate only three. Hence, a possible strategy for a speaker might be to always select the most discriminatory property, given the state of the domain and the description. In this case, once a pattern is chosen, either one of the remaining two properties (colour or size) would suffice to completely distinguish the target, since their discriminatory power is equal. This strategy is embodied in a well-known computational algorithm for the automatic generation of referring expressions, the Greedy heuristic (Dale, 1989). In the psycholinguistic community, it has been proposed most explicitly in the theoretical work of Olson (1970). Olson suggested that speakers 'specify the object to the level required by the listener to differentiate the intended referent from the alternatives' (p. 244-5). One way of interpreting this, assuming an incremental procedure, is that the speaker weighs the contribution of each available property to the ultimate goal of identification, choosing the one that is most likely to help in achieving it, as the Greedy heuristic does. [1]

In contrast to these models, experimental work has suggested that speakers' content selection processes tend to rely on heuristics related to the inherent salience of certain properties. The primary source of evidence for this is that some properties – notably an object's colour – tend to be used even when they do not contribute to the identification of a target, leading to an overspecified description. By contrast, other properties, such as size (or, presumably, pattern in the sense being used here), tend to be used only when absolutely required. In the case of size, its relatively dispreferred status is likely due to its being a relative property, requiring comparison to other objects in the domain.

These results are extremely robust (see Pechmann, 1989; Belke & Meyer, 2002; Engelhardt, Bailey, & Ferreira, 2006, among many others) and appear to persist even when the colour of an object doesn't differ too starkly from that of its distractors (e.g. the target is light green, whereas a distractor is a darker shade of green; see Viethen, Goudbeek, & Krahmer, 2012). According to Pechmann (1989), this can be explained with reference to the fact that when speakers incrementally select properties, they initiate their descriptions before having completely scanned a domain. The preference for a property like colour – which may be related to its being an inherently salient attribute of perceived objects (e.g. Pechmann, 1989) – would make it more likely for that property to be selected before others.

In short, the evidence suggests that a property's discriminatory value is not the only consideration in content selection. In the computational literature, this evidence inspired the development of the well-known Incremental Algorithm for the generation of referring expressions (Dale & Reiter, 1995). In contrast to the Greedy heuristic, the Incremental procedure works by selecting properties one by one on the basis of their preference rather than their discriminatory value. Given an ordering of properties by their preference, the algorithm considers each in turn. If a property excludes some distractors, it is included in the description, and the distractor set is updated, before considering the next property. Like the Greedy heuristic, this algorithm terminates when the description is fully distinguishing.

In our example domain, the Incremental Algorithm would thus start with the target's colour rather than its pattern. This excludes all the blue objects, leaving two other green objects. If, in the predefined preference ordering, pattern follows colour, this is the next property that would be considered. Since pattern excludes both remaining distractors, the description generated is *the green bottle with a diamond*.

Thus, there are two potentially conflicting motivations underlying content selection: discriminatory power and preference. The potential trade-off is exemplified in Figure 1, where the most discriminatory property (pattern) is not the most preferred one (colour).

In spite of the evidence for preferences stemming from overspecification, there is to our knowledge no research that explicitly tests the predictions of the two models, although some of the implications of the two strategies are evident in recent computational psycholinguistic work. Gatt, van Gompel, Krahmer, and van Deemter (2011) and van Gompel, Gatt, Krahmer, and van Deemter (2012) propose a non-deterministic model of reference production called PRO, which follows one of two different paths, each of which involves the throw of a dice, loaded to reflect the degree of preference of a set of properties. Path 1 is only followed if there exists a property that rules out all distractors (the limiting case of what

[1] It is worth noting that the Greedy heuristic is not the only model that seeks to account for Olson's theoretical stance. A possible alternative is not to proceed incrementally, but to compare entire descriptions of increasing length, starting from those consisting of a single property, until the target is distinguishing. This would ensure a description that contains no more information than is absolutely required, something the Greedy heuristic can in fact only approximate. However, this 'Full Brevity' model, also discussed by Dale (1989), is unlikely to be psycholinguistically realistic, for three reasons: (i) it is computationally extremely expensive, since it potentially involves search through all available combinations of properties (Reiter, 1990); (ii) speakers tend to overspecify, as we discuss below; (iii) an implementation of Dale's Full Brevity model has been shown to produce output that does not match that of human speakers, compared to algorithms that are incremental in nature (van Deemter, Gatt, van der Sluis, & Power, 2012).

we have called discriminatory power): the output of the algorithm in this case is a description containing this one property. Should several properties rule out all distractors then the (preference-loaded) dice is thrown to choose one of them. Path 2 is followed if no such property exists (as in Figure 1). Here, properties are added incrementally to the description until all distractors have been removed. Which property is chosen next is based on a throw of the dice. Once all distractors have been removed, however, the dice is thrown again to determine whether to terminate or include one more property; if the latter decision is taken, then the dice is thrown again to decide whether to terminate after that, or continue, and so on.Thus, preference does not only govern the choice between properties, it also governs the decision whether or not to over-specify.

Although PRO was found to have an excellent fit to production data, it was compared to human-produced descriptions in very simple domains in which there were only two properties available (colour and size) and one property always sufficed to distinguish the target referent. Thus, it is an open question whether speakers computed relative discriminatory power, or more simply based their strategy on the limiting case, namely, the availability of a fully distinguishing property.

More recently, Frank and Goodman (2012) proposed a Bayesian model to predict property choice[2] in very simple language games in which a speaker has to choose one property (such as *blue* vs. *circle*) to describe an object in a domain. Although this work does not explicitly address identification, it is nevertheless highly relevant to the present discussion. In this model, the speaker's choice of a property given a referent is based on utility. Letting p be some property of the referent, P the set of available properties, and $|p|$ stand for the number of objects of which p is true, the likelihood of using p is

$$\frac{|p|^{-1}}{\sum_{q \in P} |q|^{-1}}$$

This definition approaches the notion of discriminatory power being discussed here, because the utility of p increases the fewer objects it is true of (i.e. the more distractors it eliminates). However, this model does not consider preference. A more serious shortcoming is that the model assumes that a speaker can only refer using a single property; thus, it would never overspecify. Indeed, it turns out that the utility function over-estimates speakers' tendency to underspecify. Consider a case where a referent is both large and green. Assume that there is an additional green distractor, but no other large distractors (size is fully distinguishing). In this case, the probability of using the property *large* works out to 0.67, with a probability of 0.33 for *green*. In the experiment reported by Gatt et al. (2011), which contained a condition precisely analogous to this one, speakers produced size-only descriptions only 17% of the time, and overspecified with both size and colour 83% of the time, a finding that tallies with figures in the literature on overspecification. Speakers never produced an underspecified description with colour only. Thus, the model of Frank and Goodman too does not satisfactorily account for the interplay between discriminatory power and preference.

In summary, the question addressed by the present paper is: To what extent do preferences trump discrimination in the process of selecting properties incrementally? We investigate the issue experimentally, using domains such as the one exemplified above. If speakers tend to prioritise properties by discriminatory power, then a property should be more often included if it is the most discriminatory one available, than if it is not. By contrast, if speakers prioritise properties by preference, then more preferred properties should be included more often than less salient properties.

The experiment

The experiment traded off the discriminatory power of properties against their degree of preference, which was determined on the basis of previous empirical work. Our aim was to investigate which of the two heuristics outlined in the preceding discussions – one that prioritises properties based on preference, or one that does so based on discriminatory power – best matches speakers' content selection strategies. If preferences are more important, then the frequency with which properties are used should be independent of how discriminatory they are in different conditions. By contrast, if discriminatory power is more important, then a property should be used more often in case it is more discriminatory, regardless of whether it is highly preferred (as colour is) or not.

Participants

The experiment was conducted at the Tilburg center for Cognition and Communication. 72 native speakers of Dutch (49 female, 23 male), all undergraduate students at Tilburg University, participated in return for course credit. All had normal or corrected-to-normal vision and none reported any problems with colour perception.

Materials and design

The experimental stimuli consisted of 36 items selected from a version of the Snodgrass and Vanderwart set of line drawings with colour and texture (Rossion & Pourtois, 2004). The items were selected on the basis of a pretest in which seven native speakers of Dutch were asked to name greyscale versions of the pictures. For the items, we selected only those pictures for which at least 5 out of the 7 speakers agreed on the name of the object. The pictures were subsequently manipulated to create a

[2]Frank and Goodman (2012)'s discussion employs the term *word* rather than *property*; however, little hinges on the difference for present purposes.

version of each in two different sizes (large/small) and four different colours (red, blue, green and grey), with three superimposed patterns (a circle, a diamond or a square).

The rationale for using these three properties was as follows. First, there is a lot of previous work indicating that colour is highly preferred over size (see above). Second, the choice of pattern as the third property was based on its having to be realised (in Dutch, the language of the experiment) as a post-modifier, while size and colour tend to be realised as pre-modifiers, with a relatively fixed order (see e.g. Gatt et al., 2011, for previous work manipulating colour and size with similar materials). To the extent that the syntactic linearisation of properties reflects their order of selection, this would suggest that pattern would be selected after the other two. Be that as it may, however, we wanted to ensure minimal variation in syntactic ordering of the properties involved.

For each item, a visual domain was constructed, consisting of a target referent indicated by a black border, and five distractors. In each domain, all objects (target and distractors) were of the same type (e.g. all were bottles). In every domain, the target could be minimally distinguished from its distractors via *any subset of two* of its properties. As an example, the bottle in Figure 1 can be distinguished from its distractors by using its colour and pattern (*the green bottle with a diamond*), its colour and size (*the large green bottle*) or its size and pattern (*the large bottle with a diamond*). Each item was used in one of three conditions; the difference between conditions lay in which property of the target had the highest discriminatory value. One property was designated the most discriminatory property (hereafter **mdp**): this property excluded four of the distractors. The other two properties were equally discriminatory and each excluded three distractors. For example, in Figure 1, the MDP is pattern.

Note that, regardless of which property was the MDP, two properties were always minimally needed to distinguish the target. A description which mentioned all three properties would be overspecified, while one that mentioned only one property would be underspecified. As a result, there is no length confound: distinguishing descriptions are equally long in all conditions, unless they are over- or underspecified. This setup excludes another possibility, namely that speakers are biased to select a single, fully distinguishing property if one exists. This could happen, for example, because when a target has such a property (e.g. the target is the only red object), it becomes so salient that it induces a 'pop-out' effect (Treisman & Gelade, 1980). While such effects have been reported in experiments on visual search, they have recently also been found to influence reference production as well (Gatt, van Gompel, Krahmer, & van Deemter, 2012). Here we are interested in testing a subtler notion of discriminatory power, in a more complex domain configuration.

In each trial, objects were presented in a sparse grid. For each item, the position of the target was fixed in advance and was the same in all conditions. Both items and participants were randomly divided into 3 groups. Item and participant groups were rotated through a Latin square so that each item appeared in each condition and each participant saw all conditions, but each participant saw each item only once. The experiment consisted of 36 trials, with 108 fillers. 36 of these were constructed using the objects with the same three properties as those used in the experiment. However, the type of a target sufficed to distinguish the target. The remaining 72 fillers consisted of targets that could be distinguished from their distractors using a variety of properties (such as stripes, spots etc).

Procedure

The experiment was run using DMDX (Forster & Forster, 2003), and used a director-matcher paradigm. Participants were divided into 36 pairs, with one randomly assigned to the role of speaker/director and the other to the role of listener/matcher. Participants did not switch roles. The director and matcher faced each other; each had a computer screen that could not be seen by the other. The speaker used a keyboard to request an item, whereupon she identified the target for the listener, who clicked on the target on his own screen. Participants were instructed to keep the interaction to a minimum, with the listener only responding by indicating to the speaker that he had finished identifying the target. The speaker's descriptions were recorded through a headset.

Data coding

Descriptions were transcribed and coded according to which of the three properties of a target (colour, size and/or pattern) were mentioned. This classification ignored the mention of the object's type (e.g. 'bottle'), which we assumed would be included in any case and which, in our design, had no discriminatory value. A description was further classified as follows (i) *Well-specified* if it contained exactly two properties (excluding type) of the target; *Overspecified* if it mentioned all three properties; or (iii) *Underspecified* if it mentioned only one property, or only the object's type. The descriptions were further coded according to whether they included the MDP or not.

Results

In what follows, we report results from logit mixed effects (LME) analyses with Condition as fixed effect and random intercepts for participants and items.

Table 1 displays the proportion of overspecified, well-specified and underspecified descriptions in each condi-

MDP	*Well-spec*	*Overspec*	*Underspec*
Colour	0.699	0.296	0.005
Size	0.685	0.310	0.005
Pattern	0.676	0.324	0.000

Table 1: Proportion of well-specified, overspecified and underspecified descriptions, by condition

tion. The number of underspecified descriptions was minimal overall (4 in total) and the rate of overspecification does not appear to differ greatly across conditions. We recoded descriptions to indicate whether each one was overspecified or not. There was no effect of condition on the likelihood of overspecification ($Z = 1.02, p > .3$), that is, the likelihood of overspecifying did not depend on which property was the MDP.

MDP	*Colour*	*Size*	*Pattern*
Colour	0.99	0.73	0.57
Size	0.99	0.76	0.55
Pattern	0.99	0.73	0.59

Table 2: Proportion of descriptions containing colour, size and pattern (columns) in each condition (rows)

Table 2 displays the proportion of descriptions containing colour, size or pattern, in each condition. There are two striking facts about the data: (i) the frequency with which any of the three properties was used was largely independent of the condition, that is, whether that property was the most discriminatory one or not; (ii) there is clear evidence for preferences, with colour being used more frequently than size, and size more frequently than pattern.

An LME analysis showed a highly significant effect of condition on the likelihood with which participants used the MDP ($Z = -4.33, p < .001$); this effect was also found when the analysis was repeated focusing only on well-specified descriptions, that is, those containing two of the three properties ($Z = 4.38, p < .001$). The results show quite unambiguously that whether or not the MDP was used turned out to depend on whether it was colour, size or pattern. This is a clear indication that preference trumps discriminatory power, not the other way round.

Discussion

Our results suggest that speakers are insensitive to subtle differences in the discriminatory power of properties, relying on preference-based heuristics. Previous work has gleaned evidence for such heuristics from overspecification data, which is further used to argue against the notion that speakers observe a strict interpretation of the Gricean Maxim of Quantity. The present experiment manipulated both property preference (by contrasting colour, size, pattern) and discriminatory power (by orthogonally manipulating whether each of these properties is most discriminatory).

The evidence shows that preference has an effect on how frequently a property is chosen, but speakers are relatively insensitive to subtle differences in discriminatory power. We draw this conclusion from the clear tendency to make selections on the basis of which property is involved, rather than its contrastive value. Thus, colour, for example, is highly preferred and tends to be chosen irrespective of its discriminatory power. It is possible that the marked preference for colour is due to the fact that in our domains (e.g. 1), it becomes more salient since it characterises the entire object (e.g. a bottle is green in its entirety), whereas pattern, for example, does not. However, the consistency of our results in this regard with previous work (e.g. Pechmann, 1989; Belke & Meyer, 2002) suggests that the colour preference reflects a more general tendency.

The findings directly contradict computational models such as the Greedy heuristic (Dale, 1989), as well as proposals in the psycholinguistic literature based on theoretical work such as that of Olson (1970). In contrast, it suggests that models such as the Incremental Algorithm (Dale & Reiter, 1995) are on the right track, insofar as they make choices based on preference. On the other hand, recent work suggests that this algorithm does not give a complete picture of human reference production either. One of its limitations is that it only selects a property if it excludes some of the (remaining) distractors at a given point in the procedure, something that has been shown not to hold of human speakers (Viethen, Dale, & Guhe, 2011). Another is that the procedure is entirely deterministic and always produces the same output for a given input and a given preference ordering among available properties. In contrast, human speakers appear to treat the notion of preference stochastically, so that a model that interprets preferences in terms of a probability distribution fits human data better (Gatt et al., 2011).

This brings us to our earlier discussion of probabilistic models. One interesting question is raised by the PRO model of van Gompel et al. (2012). This model first tries to find a property which fully distinguishes the target referent. Additional content selection decisions are carried out probabilistically based on preference. As we have discussed, this model has been shown to have a remarkably good fit to data elicited from human speakers, albeit in much simpler domains than the ones used here. Now, a possible generalisation of this model would be one that first looks for the most discriminatory property available, rather than a fully distinguishing one. The results of the present experiment, which explicitly excluded the possibility of there being a single distinguishing property for the target, suggest that this would not improve its goodness of fit. However, it should be noted that our results are based on domains in which the difference in discriminatory power between the most distinguishing property and the others is exactly 1. Would a greater

difference motivate speakers to select the MDP, even if it was highly dispreferred? A positive answer to this question would imply that the PRO model can indeed be generalised to look for highly discriminatory properties, but only if their discriminatory value was relatively high, making them very visually salient. Thus, sensitivity to discriminatory power might fall on a continuum.

A similar point can be made about Frank and Goodman (2012)'s Bayesian model, which estimates the likelihood of a property being used for a referent as a function of the number of potential referents of that property, and the number of properties that the referent may be distinguished by. Modulo the simplification inherent in this work, namely that referents are to be distinguished using a single property, it would be interesting to investigate to what extent this notion of utility is also gradable and impacts visual salience.

Conclusions and future work

This paper investigated content selection in reference production. It addressed the possible trade-off between (i) the discriminatory power of a property, that is, the extent to which it is likely to help in the task of distinguishing a referent from its distractors and (ii) the extent to which a property is preferred. Our results suggest that subtle differences in discriminatory power do not influence content selection choices. One question that is left open by the present work is whether larger discriminatory power differences would alter these findings, something that we plan to investigate in future work.

Acknowledgments

Albert Gatt and Emiel Krahmer were supported by a grant from the Netherlands Organization for Scientific Research (NWO) to the project *Bridging the gap between psycholinguistics and computational linguistics: The case of Referring Expressions*. Thanks to Jette Viethen and Ruud Koolen for useful comments on a previous version of this paper.

References

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

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*(2), 237–266.

Dale, R. (1989). Cooking up referring expressions. In *Proc.* ACL'89.

Dale, R., & Reiter, E. (1995). Computational interpretation of the Gricean maxims in the generation of referring expressions. *Cognitive Science*, *19*(8), 233–263.

Engelhardt, P. E., Bailey, K., & Ferreira, F. (2006). Do speakers and listeners observe the Gricean Maxim of Quantity? *Journal of Memory and Language*, *54*, 554–573.

Forster, K. I., & Forster, J. C. (2003). DMDX: A windows display program with millisecond accuracy. *Behavior Research Methods, Instruments, & Computers*, *35*, 116-124.

Frank, M., & Goodman, N. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*, 998.

Gatt, A., van Gompel, R., Krahmer, E., & van Deemter, K. (2011). Non-deterministic attribute selection in reference production. In *Proc. PreCogSci'11.*

Gatt, A., van Gompel, R., Krahmer, E., & van Deemter, K. (2012). Does domain size impact speech onset time during reference production? In *Proc. CogSci'12.*

Grice, H. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and semantics: Speech acts.* (Vol. III). New York: Academic Press.

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

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

Olson, D. R. (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.

Reiter, E. (1990). The computational complexity of avoiding conversational implicatures. In *Proc.* ACL *'89.*

Rossion, B., & Pourtois, G. (2004). Revisiting Snodgrass and Vanderwart's object databank : The role of surface detail in basic level object recognition. *Perception*, *33*, 217–236.

Searle, J. R. (1969). *Speech acts: An essay in the philosophy of language.* Oxford: Oxford University Press.

Treisman, A., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136.

van Deemter, K., Gatt, A., van der Sluis, I., & Power, R. (2012). Generation of referring expressions: Assessing the Incremental Algorithm. *Cognitive Science*, *36*(5), 799-836.

van Gompel, R., Gatt, A., Krahmer, E., & van Deemter, K. (2012). PRO: A computational model of referential overspecification. In *Proc.* AMLAP *'12.* Trento, Italy.

Viethen, J., Dale, R., & Guhe, M. (2011). Serial dependency: Is it a characteristic of human referring expression generation? In *Proc. PreCogSci'11.*

Viethen, J., Goudbeek, M., & Krahmer, E. (2012). The impact of colour difference and colour codability on reference production. In *Proc. CogSci'12.*