Reference Production as Search

Gatt, Albert; Krahmer, Emiel; van Deemter, Kees; van Gompel, Roger P G

Published in:
Cognitive Science

DOI:
10.1111/cogs.12375

Publication date:
2016

Document Version
Peer reviewed version

Reference production as search: The impact of domain size on the production of distinguishing descriptions

Albert Gatt

Institute of Linguistics, University of Malta

Tilburg center for Cognition and Communication (TiCC), Tilburg University

Emiel Krahmer

Tilburg center for Cognition and Communication (TiCC), Tilburg University

Kees van Deemter

Department of Computing Science, University of Aberdeen

Roger P.G. van Gompel

School of Psychology, University of Dundee

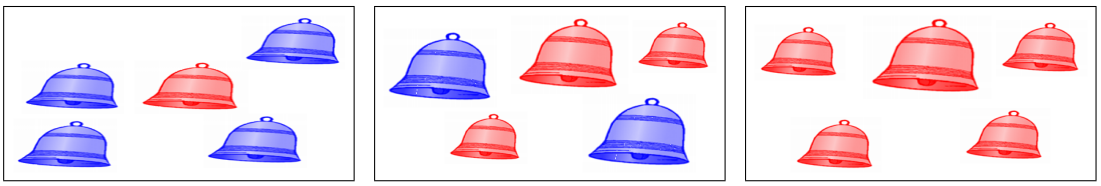
Abstract

When producing a description of a target referent in a visual context, speakers need to choose a set of properties that distinguish it from its distractors. Computational models of language production/generation usually model this as a search process and predict that the time taken will increase both with the number of distractors in a scene and with the number of properties required to distinguish the target. These predictions are reminiscent of classic findings in visual search; however, unlike models of reference production, visual search models also predict that search can become very efficient under certain conditions, something that reference production models do not consider. This paper investigates the predictions of these models empirically. In two experiments, we show that the time taken to plan a referring expression – as reflected by speech onset latencies – is influenced by distractor set size and by the number of properties required, but this crucially depends on the discriminability of the properties under consideration. We discuss the implications for current models of reference production and recent work on the role of salience in visual search.

The authors would like to thank the anonymous reviewers of this manuscript for many insightful comments and discussion. AG and EK received financial support from The Netherlands Organization for Scientific Research, via a Vici Grant (NWO Grant 27770007), which is gratefully acknowledged. KvD was supported by the RefNet project (EPSRC award EP/J019615/1). Thanks are also due to Ms Manon Yassa and ms Kristel Bartels for their help in conducting the experiments and transcription of the data.

Introduction

Reference to objects in visual scenes is pervasive in everyday communication. An intended referent, or target, is typically identified through the production of a description in which the speaker includes a subset of the properties of the target referent she has in mind. In other words, the speaker needs to perform *content determination* to establish the properties to mention in a description of her intended referent, thereby enabling the listener to identify it. In Figure 1(a) below, for example, it is immediately obvious that the object in the centre can be distinguished on the basis of its colour (*the red bell*); by contrast, colour alone won't do the trick in Figure 1(b), and it is arguably redundant in Figure 1(c), though speakers might include it anyway (producing *the large red bell*).



(a) A *red bell* among blue distractors

(b) A *large red bell* among large blue and small red distractors

(c) A *large bell* among smaller distractors

Figure 1. Visual domains with different combinations of identifying features for a target.

Content determination has often been modelled as a *search* through the target's properties and combinations thereof (cf. Bohnet & Dale, 2005). At the heart of a search-based algorithm is the abstract concept of a 'state'. For example, a REG algorithm starts from an initial state consisting of an empty description and the number of distractors that still need to be excluded. The goal state is one in which the description contains a set of properties which jointly distinguish the target from its distractors (thus, the

description is non-empty, and the set of remaining distractors is empty). A typical search algorithm recursively expands its current state into a set of possible subsequent states. For example, the initial state in REG could be expanded into subsequent possible states in which the description contains one property (size, colour, the type of the referent, etc). Which state the algorithm moves to next depends on the heuristics built into it. For example, an algorithm might choose to move to the next state (i.e. add a property to the description) on the grounds that it is the one which is most likely to exclude the largest number of the distractors remaining in the current state.

A number of influential computational models of Referring Expression Generation (REG; see Krahmer & van Deemter, 2012, for a review) are based on this view; these are summarised in an algorithm schema discussed in the next section. But if speakers do perform search in the manner informally sketched above, then one of the things that is likely to influence the speed of this process is the number of distractors against which a target has to be compared, since this is necessary to determine whether a property will help in achieving a distinguishing description. On the other hand, situations could be envisaged in which this search is conducted more efficiently by a human being than this search metaphor would suggest. To take an example, trying to verbally distinguish one person in a crowd might be a slow and time-consuming process, unless that person happens to be the only one wearing a very contrastive colour of clothing, in which case, determining whether that particular property will help probably wouldn't need an exhaustive comparison against the distractor set.

These intuitions suggest an analogy between content determination in reference

production and findings in the visual search literature, where participants are given a description of such a target and need to scan a visual scene to verify its presence or absence. As we shall see, the search literature has shed light on a range of difficulty in search, from very fast identification times for certain targets, which are unaffected by the number of distractors, to cases where search is slowed down as the distractor set gets larger.

The question we address in this paper is whether production latencies are impacted by distractor set size and to what extent this also depends on the type and number of properties required to identify the referent in a given domain. In the experiments reported below, we focus on speech onset time, that is, the time taken to initiate an identifying description of a target, as an indicator of the time speakers spend *planning* the content of a description prior to initiating an utterance.

Our starting point in addressing this question is the predictions of some computational models of content determination for reference production. While the models we investigate do not typically aim to make predictions about human reference production, they have often been motivated by psycholinguistic findings. Indeed, it has been argued that such models can be leveraged to test predictions about the cognitive processes underlying production (van Deemter, Gatt, van Gompel, & Krahmer, 2012). We propose to view these as *process* models. In cognitive modelling, such models are distinguished from *product* models (e.g. Vicente & Wang, 1998; Sun, 2008; Lewandowsky & Farrell, 2010) in that they aim to model the manner in which a given function is performed. In contrast, product models focus on the relation between inputs to a system (e.g., a

domain and an intended referent) and the outputs that it generates (e.g., a referring expression), without making any claims about the manner in which that mapping comes about.

Visual properties and distractors in reference production and visual search

The relationship between language processing and visual context is a central problem for psycholinguistics, which has been given new impetus over the past two decades, for example, through eye-movement studies in the visual world paradigm (e.g. Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995; Altmann & Kamide, 1999; Chambers, 2002; Knoeferle, Crocker, Scheepers, & Pickering, 2005; Griffin & Bock, 2000; Brown-Schmidt & Tanenhaus, 2006, *inter alia*).

A related body of work within Artificial Intelligence has also explored the relationship between vision and language, a theme that has also become dominant in research on Natural Language Generation (NLG Reiter & Dale, 2000), which focusses on the design of systems that generate text or speech in natural language from non-linguistic data. Broadly, NLG systems have addressed the relationship between vision and language in two different ways. One concern has been to automatically generate coherent, 'high-level' descriptions of visual scenes or images (e.g. Farhadi et al., 2010; Elliott & Keller, 2013; Kulkarni et al., 2013; Yatskar, Galley, Vanderwende, & Zettlemoyer, 2014). A different body of work has focussed on generating identifying descriptions of specific objects within a scene, that is, on content determination during Referring Expression

Generation (REG; e.g. Appelt, 1985; Dale, 1989; J. Kelleher, Costello, & Van Genabith, 2005; Stoia & Shockley, 2006; Campana, Tanenhaus, Allen, & Remington, 2010; Krahmer & van Deemter, 2012; J. Kelleher et al., 2005; Stoia & Shockley, 2006; Campana et al., 2010; Garoufi & Koller, 2013; Kazemzadeh, Ordonez, Matten, & Berg, 2014).

Our focus in this paper is on content determination for definite descriptions, to identify targets in visual domains of the sort depicted in Figure 1, with varying distractor set sizes.[1] Content determination in REG parallels the conceptualisation process in Levelt's well-established model of speech production, in which it precedes grammatical formulation and articulation (Levelt, 1989, 1999).

An influential class of REG models, based on the work of (Dale & Reiter, 1995), have modelled the property selection process 'incrementally' (cf. Levelt, 1989). The basic assumption is that if a speaker mentions, for example, that an object is red, she implies that at least some of the distractors are not red. Thus, such algorithms assume that the properties which are mentioned have a contrastive function. Put somewhat differently, these algorithms assume that the path from the initial to the goal state consists of intermediate states, at each of which the addition of a property reduces the remaining set of distractors, thereby bringing the algorithm closer to the goal at each step. Algorithmically, this assumption is incorporated in a procedure whereby a

[1] In what follows, we consider only properties such as size or colour, excluding any consideration of the object's type or category (e.g. *bell*). This is partly following proposals in the REG literature (e.g. Dale & Reiter, 1995), and partly because in the visual domains used for our experiments, target referent and distractor objects are all of the same type. Note, however, that this exclusion does not affect the predictions made by computational models on the nature of search during reference production.

1: **while** not all distractors have been ruled out **do**
2: select a new property p of r
3: **if** p excludes some distractors **then**
4: add p to the D under construction, and
5: remove the distractors ruled out by p
6: **end if**
7: **end while**

Figure 2. A general, 'incremental' content determination procedure.

candidate property – say, *red* in Figure 1(a) – is checked and if it removes at least some distractors, then the property is included in the description. Taking r to be a target referent and D the description under construction for r (that is: D is a set of properties, initially empty), then the underlying algorithm that these models implement can be schematised as in Figure 2.

In words: as long as there are still distractors that are not ruled out (line 1), the model checks, for each new property of the target (2), whether it helps to rule out any remaining distractors (3); if so, this property is added to the description under construction (4) and the distractor set is updated (5).

It is worth emphasising that the model outlined here is an abstraction of various existing models, which seeks to bring out some of their common assumptions. Differences between various well-known content determination models are primarily due to the way in which a new property becomes a candidate for inclusion (i.e., with how line 2 is

actually implemented). An algorithm may prioritise certain properties over others for a lot of different reasons, such as their perceptual salience, their frequency in a corpus, how discriminatory they are (that is, how many distractors they exclude) and so on. To take an example, Dale (1989) proposed a Greedy heuristic that tries out properties in order of their discriminatory power: properties that rule out many distractors are preferred over properties that rule out only a few. Thus, the algorithm first determines which property rules out most distractors, and then incrementally extends the description based on which property has most discriminatory power at that stage. In a related vein, recent stochastic accounts (e.g. Frank & Goodman, 2012) model the production process as a search for the most informative property available to identify a target referent.

A different heuristic is incorporated in the Incremental Algorithm (Dale & Reiter, 1995), as well as recent stochastic versions thereof (e.g. Mitchell, van Deemter, & Reiter, 2013). This model assumes the existence of a preference order; as a result, properties are considered for inclusion in a description in order of their 'preference'. Applied to visual domains, this strategy is partly inspired by psycholinguistic work showing that certain properties are preferred by speakers, who are more likely to include them in a description even if they are not absolutely required for identification. A robust finding is that colour tends to be included in this way much more than size (Pechmann, 1989; Deutsch & Pechmann, 1982; Eikmeyer & Ahlsèn, 1996; Koolen, Gatt, Goudbeek, & Krahmer, 2011; Engelhardt, Bailey, & Ferreira, 2006; Belke, 2006; Arts, 2004). Thus, a speaker is more likely to refer to the target in Figure 1(c) as *the large red bell* than she is to refer to the one in Figure 1(a) as *the small red bell.* Both of these descriptions are

overspecified, in the sense that they contain a property that isn't strictly required for identification (though such redundancy is known to serve other communicative purposes; see for example Jordan & Walker, 2005). The fact that more overspecification occurs with colour than with size suggests a 'preference' for the former.

In visual domains such as these, the Incremental Algorithm has often been implemented to check properties such as colour before those considered as prototypically gradable, such as size, on the grounds that the latter is dispreferred because determining the size of a target requires comparison to the distractors. However, it turns out that the dichotomy between 'crisp' and 'gradable' is not so clear-cut and the likelihood of selection of a property depends on how contrastive or discriminable it is in the context of a scene (Viethen, Goudbeek, & Krahmer, 2012; van Gompel, Gatt, Krahmer, & van Deemter, 2014), as well as how diagnostic of the object under consideration (Sedivy, 2003; Westerbeek, Koolen, & Maes, 2015).

REG models and search efficiency

One prediction that the model in Figure 2 makes is related to its treatment of distractors. This can be made explicit in relation to the procedure's worst-case complexity, that is, the function that serves as an upper bound for the time taken by the procedure to identify a distinguishing set of properties to include in D.[2] The computational complexity will differ depending on the way in which the general procedure is implemented. The crucial parts of the algorithm are lines 2 and 3 in Figure 2.

[2]Our discussion of complexity partly follows the exposition in Dale and Reiter (1995), with some modifications.

In the Greedy Algorithm interpretation (Dale, 1989), every time a property is considered (line 2), it needs to be evaluated for its discriminatory power against all the other remaining properties. This involves checking how many distractors the property rules out, that is, checking for each distractor whether the property applies to it or not. Since in the model every distractor has to be checked to see if it has the property, this is usually assumed to be a serial process, an assumption common to all models under discussion here.

Suppose there are n_p properties available, and n_d distractors in the domain. Then, at each iteration, the Greedy algorithm needs to check at most n_p properties against n_d distractors, resulting in complexity $O(n_p n_d)$. In this way, this procedure compares properties against each other to identify the most discriminatory one at each iteration. By contrast, in the Incremental Algorithm interpretation, the order with which properties are checked is fixed in advance by the preference order, obviating the need to make comparisons between properties. Nevertheless, this procedure still needs to check whether a candidate property has any discriminatory value against the remaining distractors, making $O(n_d)$ comparisons at most, at each iteration in line 3.

Note that, in both of these cases – indeed, in any instance of an algorithm that fits the outline in Figure 2 – the number of distractors plays a role in determining the amount of 'effort' expended on finding a distinguishing description, because every candidate property is checked against the distractor set. The general model therefore predicts that the time taken to identify a target using a referential description will increase in the number of distractors. A further prediction is related to the number of

properties that are eventually included in D. Suppose D is of length n_l. For example, $n_l = 2$ for the description *the large red bell* for the target in Figure 1(b). This means that the procedure will have made n_l iterations, each time conducting a serial check against the distractor set (line 3) to include one property (at line 4). In summary, two predictions stem from incremental models of reference production:

1. The time taken to produce a description increases with the number of distractors in the domain;

2. The time taken to produce a description also increases with the number of properties in the description.

Crucially, these predictions are made independently of the factors which are known to modulate speaker choices. Even in the case of the Incremental Algorithm, where 'preference' governs which properties are considered first, a property of a target is checked against the distractor set in the same way, irrespective of the property.

The foregoing discussion highlighted a number of reasons to question this. Recall that research on reference production suggests that speakers use certain properties, such as colour, with greater likelihood than others, all other things being equal. In part, this may be due to the centrality of colour in object representations, a proposal made early on by Pechmann (1989) to explain his overspecification results, and which receives some support from research on the central role of colour in object recognition (e.g. Wurm, Legge, Isenberg, & Luebker, 1993; Naor-Raz, Tarr, & Kersten, 2003). On the other hand, other research has argued for preferences in reference production as arising from early (that is, pre-linguistic) perceptual processes (Belke & Meyer, 2002; Belke, 2006).

Under this account, a preference for a property is evinced if the contrast between a target referent and its distractors on that specific dimension is highly salient. Indeed, as we have seen, recent work also shows that selection of colour may become less likely in domains with more colour variation (and less likely when colour is highly diagnostic of an object category); similarly, size is more likely to be used if size differences between target and distractors are made much larger.

Given the potential effects of visual salience , not all properties may be selected as predicted by Figure 2. While properties such as size, which tend not to be very salient, may require the kind of serial, one-by-one checking of each distractor predicted by the model, properties that are more salient, as colour often is, may not. The literature on visual search tasks sheds further light on these issues.

Visual search

The standard visual search task requires participants to scan a domain and verify the presence or absence of a target (Wolfe, 2010). Research using this paradigm has often focussed on two components, namely (i) the target *template*, the representation of the target based on its features (e.g. Duncan & Humphreys, 1989), which is usually formed prior to the commencement of search on the basis of an instruction (*is there a red vertical?*); and (ii) the visual display, in which the target may or may not be present, and in which varying numbers of distractors are found.

The number of distractors is known to influence the speed with which a target can be found, under certain conditions. In a classic study, Treisman and Gelade (1980) found

that search for single features (e.g. defined by the template RED) did not depend on the size of the display, evincing a 'pop-out' effect typified by a flat RT × set size slope. By contrast, a search for conjunctions of features showed a linear increase in search time as a function of domain size. Furthermore, search time was found to exhibit a roughly 2:1 ratio between target-absent and target-present trials; this was explained on the grounds that in the target-absent case, participants had to search a display exhaustively, while they only needed to search through half of a display on average in the target-present case (cf. Nakayama & Joseph, 1998).

Feature Integration Theory (FIT Treisman & Gelade, 1980) accounted for these findings based on a two-stage model. Parallel, 'preattentive' processes sensitive to individual features account for the pop-out effects observed in single feature search. These are strongly dependent on the discriminability or salience of the features in question (Treisman & Gormican, 1988; Itti & Koch, 2001), a factor that also plays a role in many computational models whose task is to predict the salient regions in a scene (e.g. Itti, 2005; Walther & Koch, 2006; Achanta, Hemamiz, Estraday, & Süsstrunky, 2009; Erdem & Erdem, 2013). Discriminability is a relative notion. For example, search for colour-defined targets becomes more difficult if the target colour is collinear with distractor colour Bauer, Jolicoeur, and Cowan (1996). Similarly, search for a target which is distinguished by size, a feature which has also been claimed to be subject to parallel, preattentive processing (e.g. Stuart, Bossomaier, & Johnson, 1993), turns out to depend on numerous context effects that modulate its discriminability (e.g. Busch & Müller, 2004). In contrast to these bottom-up, feature-driven processes, FIT posits a role for

top-down, serial, attention-driven processes which are responsible for feature binding and hence come into play during conjunction search. Thus, FIT is a two-stage architecture, at whose core is the distinction between parallel, bottom-up and serial, top-down processes (cf. Neisser, 1967).

Subsequent work questioned the adequacy of this dichotomy in explaining the data, both on empirical and theoretical grounds. For example, a large scale meta-analysis of search data by Wolfe (1998) found no evidence of bimodality which could be taken to correspond to different search processes. Although bimodality is not a necessary and sufficient criterion for identifying distinct processes, subsequent follow-up research has nevertheless suggested a departure from the standard bottom-up/parallel versus top-down/serial dichotomy (Haslam, Porter, & Rothschild, 2001). At the same time, arguments were put forward against the split between 'pre-attentive' (bottom-up) and attentive processes (Nakayama & Joseph, 1998). Indeed, the empirical evidence supports a more nuanced view of the relationship between search time and distractor set size. Faster versus slower search is known to be affected by a variety of factors, including, among others, the extent to which a display affords the formation of perceptual groups (Nakayama & Silverman, 1986; He & Nakayama, 1995; Nakayama & Joseph, 1998; Nordfang & Wolfe, 2014); the salience of individual features that are reliably correlated with a conjunction (Wolfe, Cave, & Franzel, 1989; Sobel & Cave, 2002; Found, 1998); and the similarity between target and distractors, as well as the specificity of the template (Duncan & Humphreys, 1989; Malcolm & Henderson, 2009). Search for complex targets is also affected by the nature of the experimental paradigm (e.g. cueing vs. standard

visual search; cf. Palmer, 1994, 1995; Vickery, King, & Jiang, 2005); and by the search strategy afforded by the display (e.g. the presence of subsets; cf. Friedman-Hill & Wolfe, 1995). Guidance (Olds, Cowan, & Jolicoeur, 2000b, 2000c, 2000a) or preview of features in a target conjunction (Olds & Fockler, 2004) also facilitate search for complex targets. One set of results has shown evidence for linguistic guidance. Spivey, Tyler, Eberhard, and Tanenhaus (2001) showed that auditory presentation of a target template or description, incrementally and concurrently with the display, results in shallower RT $\times$ set size slopes (though it has been argued that this form of linguistic guidance is dependent on the speech rate with which the description of the target is delivered; Gibson, Eberhard, & Bryant, 2005). This claim is also supported by evidence of search facilitation when features of the target are incrementally presented using a visual, rather than a linguistic modality (Chiu & Spivey, 2012). Reali, Spivey, Tyler, and Terranova (2006) provided further confirmation by replicating the findings of Spivey et al. (2001), while also showing that the order in which information is delivered matters: describing the target using colour followed by orientation facilitated search more than did the opposite order. This echoes findings by Olds and Fockler (2004), who found a similar order effect using visual preview of stimuli. Crucially Reali et al. (2006)'s design excludes the possibility that linguistic guidance effects are an artefact of blocked designs, or that they are due to an odd-one-out search strategy.

An important outcome of this body of work has been a more unified view of the processes underlying visual search. Some models explicitly argue for a continuum of search efficiency determined by such factors as target-distractor similarity (Duncan &

Humphreys, 1989) and competition between multiple features across the visual field as a function of both top-down and bottom-up processes, leading to bias in attentional allocation (Desimone & Duncan, 1995; Desimone, 1998). Simulations such as those made by (Reali et al., 2006) lend support to these unified models, by showing that multiple factors – visual features and linguistic input – can contribute to the gradual emergence of a region in the visual field as the likely target for attention. On the other hand, current models which maintain a two-stage architecture, such as Guided Search (GS; Wolfe et al., 1989; Wolfe, 1994, 2007), differ from the classic FIT model in that attention and selection are explicitly guided by a limited number of visual features (reviewed by Wolfe & Horowitz, 2004).

The present study

How does content determination for referring expressions relate to visual search? Our discussion of the two bodies of literature concerning these processes highlights points of convergence, but also important differences.

The nature of search in reference production could be described as *object-driven*: in a typical referential situation, the speaker has a target referent in focus, whose properties are known, or at least accessible, to her, and from among which she needs to select a distinguishing subset to enable an interlocutor to identify the same object. As the family of computational models we have discussed make explicit, this results in a search within the space defined by these properties and their combinations. Furthermore, this conceptualisation or content determination process is incremental (Pechmann, 1989;

Levelt, 1999).

By contrast, in a typical visual search paradigm, search is *template-driven*: success is defined as matching a target against a description, or concluding that no such target is present (Wolfe, 2010). Research showing that search effort can be modulated by the concurrent verbal delivery of the target description (Spivey et al., 2001; Reali et al., 2006), or by facilitation of search through the concurrent presentation of individual target features in the visual modality (Chiu & Spivey, 2012) suggests that incrementality can also play a role in this task.

Despite the differences between them, in both cases search has often been assumed to require comparison between elements of the display, under certain conditions. In the visual search case, this assumption has been made on the basis of evidence that certain types of search are made slower with increasing numbers of distractors although, as we have seen, there are many factors that modulate this, as well as competing accounts of the causes of search inefficiency. In the reference production case, the evidence is more indirect, stemming from preferences for properties observed in production data. Thus, there is an open question related to the extent to which efficiency in content determination is affected by distractor set size, and by the nature of the properties required to identify the target. The family of computational REG models discussed here make explicit predictions about this, which remain untested.

In what follows, we empirically investigate the two predictions of computational REG models outlined above in two experiments. Experiment 1 compares reference production in conditions where size alone, or both size and colour, suffice to distinguish

a target; Experiment 2 compares the situation where colour alone can do the trick, compared to the same condition where size and colour are required.

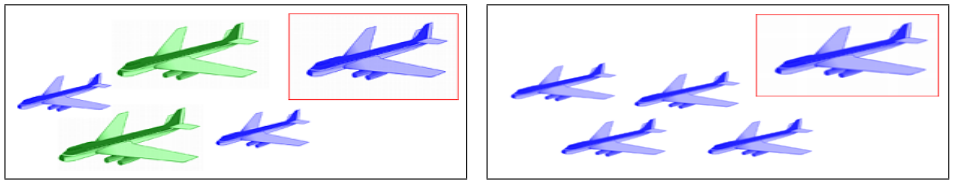
Our experiments make simplifying assumptions about the set of distractors by focussing on a fixed array of objects that constitute a referential domain. This is somewhat akin to the assumption in many of the visual search experiments described above, where the entities and features manipulated are simple and well-defined. The picture is of course far more complex in real-world scenes. As Wolfe (2010) notes, for instance, the process of verifying whether a cow is present in a field will be informed not only by the features of that cow in relation to the other elements of the scene, but also by prior knowledge of what such scenes typically consist of. Recent work on vision has begun to explicitly address how attention and search are influenced by such factors, including prior expectations about typical scene structure (see Oliva & Torralba, 2007, and references therein), semantic factors (e.g. Henderson, Brockmole, Castelhano, & Mack, 2007; Belke, Humphreys, Watson, Meyer, & Telling, 2008; Hwang, Wang, & Pomplun, 2011) and task-based factors (e.g. Einhäuser & Koch, 2008; Awh, Belopolsky, & Theeuwes, 2012). This research has culminated in models which use of such global factors to modulate the impact of low-level features in the computation of salience, in order to predict likely regions where attention will be deployed (e.g. Torralba, Oliva, Castelhano, & Henderson, 2006; Kanan, Tong, Zhang, & Cottrell, 2009).

Ignoring such 'global' or 'contextual' information is a simplifying step, but one that permits us to study the predictions of the algorithms under consideration more precisely. Nevertheless, we return to the role of contextual information in visual search

in the concluding section of this paper, where we also speculate on its implications for reference production in light of the results obtained in the present work.

Experiment 1

In our first experiment, participants were exposed to visual domains with a designated target object surrounded by a number of distractors. The target was distinguishable from the distractors either on the basis of both its size and colour (Figure 3(a)) or its size alone (Figure 3(b)). Participants had to identify the target using a spoken description. We focussed on the speech onset time for the description, that is, the time from the presentation of the visual scene to the beginning of their utterance.



(a) Colour and size required (b) Size only required

Figure 3. Two example domains with four distractors, from Experiment 1

If the predictions of current REG models are correct, the distractor set size should impact speech onset time. Thus, the domains in Figure 3, containing four distractors, speakers should be slower compared to domains where there are two, for example. Furthermore, we would expect the scenario displayed in Figure 3(a), where a conjunction of properties is required, to evince longer speech onset times compared to the single-property case.

Participants

The experiment was conducted at the Tilburg center for Cognition and Communication. Forty native speakers of Dutch participated in return for course credit.

Materials and design

The experimental stimuli consisted of 64 items selected from a version of the Snodgrass and Vanderwart (1980) 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 version of each in two different sizes (large and small) and four different colours (red, blue, green and grey). For the size manipulation, small images covered 45% of the pixel area of large images, excluding the white background.

For each item, 8 versions of a visual domain were constructed, each consisting of a target referent surrounded by a red border, and a number of distractors, as shown in Figure 3. The 8 versions represented combinations of the following two factors:

- *Properties* (2 levels): On half the trials, the target could be distinguished on the basis of size only (S), as in Figure 3(b). On the remaining trials, both colour and size (CS) were required to distinguish the target, as in Figure 3(a).

- *Distractors* (4 levels): There were 2, 4, 8 or 16 distractors in addition to the target, representing increasing domain size. Figure 3 is an example of the 4-distractor

condition.

In each domain, all objects (target and distractors) were of the same type (e.g. all were aeroplanes). In the S trials, distractors were identical to the target except for their size. Distractors were also identical to each other (e.g. the target was a small blue aeroplane and all distractors were large blue aeroplanes). In the CS trials, half the distractors were identical to the target except for their size and the other half were identical to the target except for their colour (e.g. the target was a large blue aeroplane, half the distractors were small green aeroplanes and the other half were large blue aeroplanes). Thus, distractors in the visual display in this condition fell into two subsets.

In addition to the experimental items there were 108 fillers. In 64 of these, the target could be distinguished using size only or both size and colour, as in the critical trials. However, there was variation in the types of distractors (not all distractors were of the same type as the target). In the remaining 64 fillers, the target could be distinguished by using its type only. There were equal numbers of fillers containing 2, 4, 8 or 16 distractors.

In each trial, objects were presented in a sparse grid. For each of the items, a position in the grid was randomly fixed in advance, so that a given item (such as an aeroplane) always appeared in the same position as a target, irrespetive of the condition. The position of the distractors was also fixed in the 2-, 4-, 8- and 16-distractor conditions. Both items and participants were randomly divided into 8 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 64 items were placed in a pseudo-random order at the outset, so that they occurred in exactly the same order irrespective of condition, for all participants. Among the 64 trials, there were approximately equal numbers of targets in the two different sizes and the four different colours.

Procedure

Participants did the experiment individually in a sound-proof booth, wearing a headset through which their descriptions were recorded. The experiment was run using the DMDX package for stimulus presentation (Forster & Forster, 2003). Participants were asked to imagine that they were describing objects for a listener who could see the exact same objects but did not know which one was the target referent. In order to avoid the use of descriptions containing locative expressions (e.g. *the one in the top right*), participants were also told that their putative listener would see the objects in different positions. None of the participants used locative expressions in the experiment. Participants were also instructed to speak naturally and clearly, but to respond as fast as possible given these conditions.

A trial was initiated with a warning bell and a fixation cross appearing for 500ms in the middle of the screen. Subsequently, the visual domain appeared with the target surrounded by a red border. After they had described the target, participants pressed the Enter key on their keyboard to move to the next trial.

Trials were presented in two blocks to allow participants to take a break. Speech

onset time was measured using the DMDX voice trigger from the point when the visual domain was presented to the point when a participant began to speak.

Data pre-processing

Descriptions were transcribed and annotated for whether they contained size, colour or both. Descriptions in the S condition which contained both size and colour were classified as *overspecified*. Descriptions in the S condition which contained only colour, or those in the CS condition which contained only one of the two properties, were classified as *underspecified*. All other descriptions were classifed as *minimally specified*. Data from two participants was excluded because they produced utterances which compromised the calculation of speech onset time (for example, starting all of their descriptions with *I see a...*[3]). In what follows, analyses are conducted from data from the remaining thirty-eight participants.

Table 1 displays frequencies and proportions of well-specified, overspecified and underspecified descriptions, by condition and overall. The relatively high proportion of overspecifications in the S condition is compatible with previous findings, where the rate of overspecification is typically similar or higher, as speakers tend to use colour non-contrastively. For example, Gatt, van Gompel, Krahmer, and van Deemter (2011) report between 78% and 80% redundant use of colour in conditions where size suffices to distinguish an entity; Belke (2006) report similar proportions (ca. 87%).

[3]In such cases, it is possible that content planning for the referring expression is going on during speech. This would mean that speech onset time would not reflect planning time before the utterance is initiated.

	Minimally specified	Overspecified	Underspecified
CS2	295 (97)	0	9 (3)
CS4	291 (95.7)	0	13 (4.3)
CS8	296 (97.4)	0	8 (2.6)
CS16	297 (97.7)	0	7 (2.3)
S2	145 (47.7)	157 (51.6)	2 (0.7)
S4	142 (46.7)	157 (51.6)	5 (1.6)
S8	138 (45.5)	162 (53.5)	3 (1)
S16	132 (43.56)	168 (55.4)	3 (1)
overall	1736 (71)	644 (26.5)	50 (2)

Table 1: Frequencies and percentages (in parentheses) of minimally specified, underspecified and overspecified descriptions in Experiment 1.

In what follows, we report statistical analyses based only on the minimally specified descriptions, excluding over- and underspecified cases, although we also report speech onset times for overspecified and underspecified descriptions, for ease of comparison. Our exclusive focus on minimally specified descriptions is due to the following reasons. In underspecified descriptions, participants presumably did not check against the entire distractor set to see whether a selected property combination was distinguishing, whereas when participants overspecified, the inclusion of a redundant property may not have involved such a check because it was extra information, added after initial content

planning had determined the minimal requirements to identify the referent.

Speech onset times were manually tuned using CheckVocal (Protopapas, 2007), a program for the detection and correction of voice key mistriggers (due to lip smacks, coughs, background noise etc) in DMDX result files. For each sound file, we ensured that the speech onset time was taken at the precise point where the participant's description began. In case the description included a determiner, this meant the onset of the determiner. In case a description began with a hesitation (e.g. *uhhhh het kleine rode bed* 'uhhhh the small red bed'), the onset time was the onset of the description following the initial hesitation.

Following tuning, an onset time was defined as an outlier if it exceeded the mean ±2SD in its condition. 112 data points (4%) were considered outliers by this criterion and were treated as missing.

Results

Table 2 displays mean speech onset times and standard deviations as a function of condition, as well as across conditions for minimally specified, overspecified and underspecified descriptions. Among minimally specified descriptions, there appears to be an increase in onset time in the CS compared to the S condition. Onset time also increases with the number of distractors.

Figure 4 displays the effect of increasing distractor set sizes on speech onset time for minimally specified descriptons in the S and CS conditions. In both cases, times increase, albeit with a slight decrease for 8-distractor domains in the S condition. The

		2	4	8	16	Overall
	CS	2023 (492)	2022 (507)	2106 (525)	2140 (809)	2073 (538)
Minimally specified	**S**	1872 (458)	1991 (566)	1922 (209)	2047 (473)	1955 (506)
	Overall	1972 (486)	2012 (527)	2047 (527)	2111 (572)	–
	CS	2308 (623)	2076 (524)	2019 (727)	2699 (624)	2215 (635)
Underspecified	**S**	2202 (1314)	2571 (649)	2246 (1450)	2703 (19)	2491 (647)
	overall	1977 (491)	2013 (523)	2066 (564)	2084 (554)	–
Overspecified	**S**	1972 (487)	1992 (502)	2120 (647)	1986 (480)	2019 (537)

Table 2: Mean speech onset times (in milliseconds) and standard deviations (in parentheses)for minimimally specified, overspecified and underspecified descriptions, in each condition in Experiment 1. Overspecifications occurred only in the S condition.

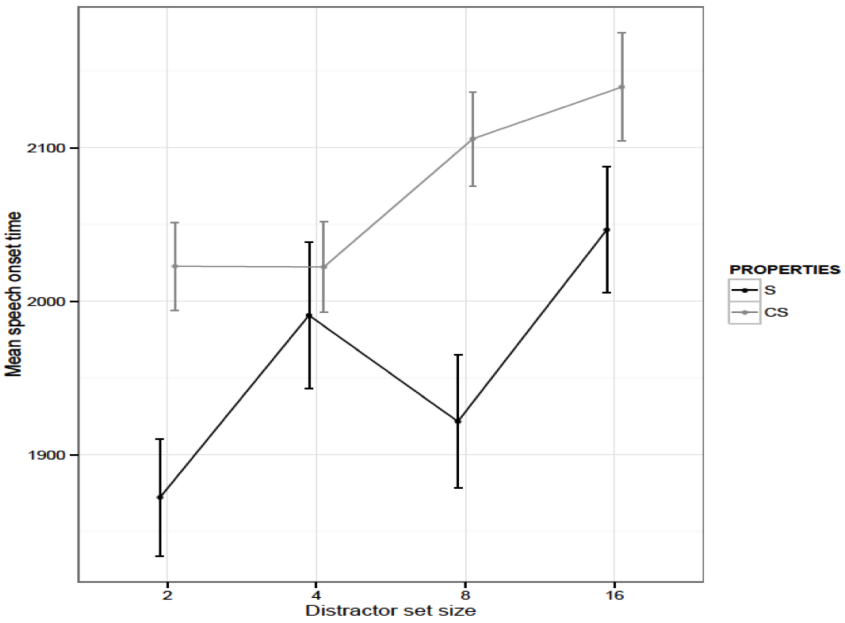


Figure 4. Mean onset times by number of distractors for minimally specified descriptons in the S and CS conditions. Error bars represent ±1 standard error of the mean

plot shows no evidence of an interaction between the two factors.

In what follows, we use Linear Mixed Effects models, reporting both model comparisons and estimates of significance of main effects and interactions.[4] We first construct a baseline (model 0) consisting only of the intercept, together with random effects. We then assess the contribution of the fixed effects of Properties and Distractors (a) separately, by constructing models incorporating each one (models 1 and 2) and comparing them to the baseline; and (b) jointly, by constructing a model with both fixed effect terms (model 3) and comparing it to the baseline. Finally (c) we compare model 3 to a maximal model incorporating the two fixed effects and their interaction (model 4). In each case, model comparison is carried out on the basis of goodness-of-fit, using Bayesian Information Criterion (BIC) and Log-likelihood estimates. Finally, we give full details of the best-fitting model, including its parameter estimates and associated significance tests.

For the purposes of the analysis, the Distractors factor was coded as numeric, since we wish to test the impact of linearly increasing numbers of distractors in the visual domain. Both the distractors and properties factors were centred to reduce collinearity and facilitate interpretation of main effects.

Following Barr, Levy, Scheepers, and Tily (2013), models were initially fitted with a maximal random effects structure, including random intercepts, and random slopes

[4]The analysis was conducted in R using the `lme4` package, version 1.1.6 (Bates, Maechler, & Bolke, 2014). Model comparisons and estimates of p-values were conducted using the `anova` and `summary` functions in the `lmerTest` package version 2.0.6 (Kuznetsova, Brockhoff, & Christensen, 2014). All significance tests are estimated from the models described below.

for both fixed effects and their interaction. Where this led to problems of convergence, we fitted the models by omitting covariances from the variance-covariance matrix; this maintains the maximal random effects structure while permitting model fitting with fewer parameters.

As a final check, the best-fitting model among the four we test is further compared to a version of the same model that includes random intercepts and slopes for item frequency, that is, the frequency of the noun for the pictures used as stimuli. This controls for possible frequency-related differences among items, which may have impacted the time taken to plan a description. Frequencies for the Dutch noun for each item were identified from the NLTenTen corpus, a Dutch web corpus of ca. 2.5 billion words constructed in 2014 and available via the SketchEngine[5]. We used log-transformed frequencies for the analysis.

Table 3 summarises the baseline model and all subsequent models, with indications of their goodness of fit. Both models 1 and 2, incorporating a single fixed effect, have a better goodness of fit than the baseline model, as reflected by the BIC and χ^2 estimates. The model that best explains the variance in the data is the one including both Properties and Distractors as main effects (model 3), as indicated by the BIC estimate. This model was significantly better than either of the models incorporating the individual fixed effects (model 1: $\chi^2 = 15.97, p < .001$; model 2: $\chi^2 = 17.26, p < .001$). As expected, given the trends displayed in Figure 4, the inclusion of an interaction term (model 4) does not improve fit over model 3, although model 4 is still significantly better than the

[5] http://www.sketchengine.co.uk

	Fixed effects	BIC	**Model χ^2**
0	Intercept only (baseline)	25312	–
1	Properties	25302	17.67^* (relative to model 0)
2	Distractors	25303	$16.37, ns$ (relative to model 0)
3	Properties + Distractors	25293	33.64^* (relative to model 0)
4	Properties × Distractors	25300	$0.04, ns$ (relative to model 3)

Table 3: Model goodness of fit statistics. Models 1, 2 and 3 are compared to the baseline (model 0) to establish the contribution of Properties and Distractors separately. Model 4 is compared to model 3 to establish the contribution of the interaction. (*) indicates significantly better goodness of fit at $p < .001$.

baseline model ($\chi^2 = 33.69, p < .001$).The best-fitting model is therefore Model 3, whose details are shown in Table 4. Note that apart from an increase in speech onset time of roughly 130ms between the one- and two-property conditions, the slope for Distractors suggests an additional 50ms in speech onset time per unit increase in the number of objects in the domain. A comparison of this model to a similar model with additional random intercepts and slopes for item frequency showed no significant difference (BIC for model 3 with item frequency: 25352; $\chi^2 = 0.03, ns$). Hence, we conclude that there was no impact of frequency on speech onset time.

Parameter	**Estimate**	**Standard Err.**	*t*-value
Intercept	2051.26	48.37	42.41*
Properties	130.18	25.02	5.20*
Distractors	49.47	11.05	4.48*

Table 4: Estimes for Model 3, the best-fitting model for the data in Experiment 1, incorporating fixed effects of Properties and Distractors. (*) indicates significance at $p < .001$.

Discussion

The main effect of Properties in this experiment suggests that speakers took longer to initiate a description when they had to describe a target using both size and colour, compared to size only. Furthermore, this occurred independently of the distractor set size effect, as shown by the lack of an interaction.

Interestingly, when speakers identified a target referent using size only, there was evidence that the number of distractors impacted the efficiency of the content determination process. This is compatible with an interpretation whereby speakers needed to compare the size of the target to that of distractors in order to determine that the target was large or small; thus, the search speakers conduct to determine the size of a target is relatively inefficient, at least for the size differences manipulated here.

These findings lend support to the predictions of reference production models, which predict that distractor set size should affect efficiency, even in the single-property case. However, in the reference production literature, size has usually been found to

be 'dispreferred' by speakers, an oft-cited reason being that it is a paradigm case of a gradable property, and hence less easily codable (Belke & Meyer, 2002; Belke, 2006), compared to properties like colour. Indeed, our own data (see Table 1) suggests that speakers often overspecified and used colour in the S condition. This raises the question of whether distractor set size would impact content determination in the same way when the required distinguishing property is highly 'preferred', in the sense that speakers tend to use it very frequently, even when it is not required. Such preference data could indicate that, all other things being equal (for example, the discriminability of the property or its typicality relative to the type of object under consideration Viethen et al., 2012; Sedivy, 2003; Westerbeek et al., 2015), speakers are able to conduct a more efficient search for such properties in the course of planning a description.

Experiment 2

Experiment 2 replicated the design of Experiment 1 and maintained the condition where two properties (colour and size) are required to distinguish a target referent. However, the single-property case this time featured a colour rather than a size contrast, as shown in Figure 5. Colours were selected to be highly distinctive, in order to provide a contrast to the previous experiment where size differences may have been less discriminable.

Participants

The experiment was conducted at the Tilburg center for Cognition and Communication. Thirty-eight native speakers of Dutch participated in return for course credit.

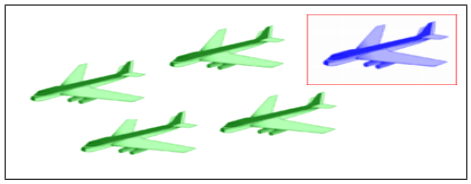


Figure 5. A domain in which colour alone distinguishes the target, from Experiment 2.

None of them had participated in Experiment 1.

Materials and design

The experimental stimuli consisted of the same 64 items used in Experiment 1, with the same size and colour values. Pictures were manipulated to create 8 versions of a visual domain representing combinations of the following two factors:

- *Properties* (2 levels): On half the experimental trials, the target could be distinguished on the basis of colour only (C), as in Figure 5. On the remaining trials, both colour and size (CS) were required to distinguish the target, as in Figure 3(a).
- *Distractors* (4 levels): There were 2, 4, 8 or 16 distractors in addition to the target, representing increasing domain size. Figure 3 is an example of the 4-distractor condition.

Once again, a pseudo-random ordering of items was set in advance. Items and participants were divided into eight groups, as before, and were rotated through a latin square to ensure a within-participants and within-items design. The different colours and sizes were again used an equal number of times for target referents across the 64 items.

Procedure

The procedure was identical to that followed in Experiment 1.

Data pre-processing

The data from three out of the thirty-eight participants had to be omitted: in two cases, participants were pressing the Enter key to move on to the next scene before having finished describing the target, resulting in incomplete recordings; a third participant underspecified more than 50% of the time.

Speech onset times were once again tuned using CheckVocal. Speech onset times that lay outside the range of the condition mean $\pm 2SD$ were classified as outliers and treated as missing. There were 86 (3.8%) outliers overall.

Descriptions were transcribed and annotated for whether they contained size, colour or both. Once again, they were coded as minimally specified, overspecified, or underspecified. Table 5 displays the frequencies of these description types, by condition. Note the much lower proportion of overspecified descriptions in the C conditions, compared to the S conditions in Experiment 1 (see Table 1). This is expected, given that size is usually found to be dispreferred with respect to colour, and especially in view of the relatively limited size contrasts used in the materials in the previous experiment.

Once again, we focus exclusively on the minimally specified descriptions in the analysis of results. Given the very small proportions of underspecified and overspecified descriptions, we refrain from reporting mean speech onset times for these types of descriptions.

	Minimally specified	Overspecified	Underspecified
C2	275 (98.2)	4 (1.4)	1 (0.4)
C4	273 (97.5)	5 (1.8)	2 (0.7)
C8	277 (99)	3 (1)	0
C16	278 (99.3)	2 (0.7)	0
CS2	275 (98.2)	0	5 (1.8)
CS4	277 (98.9)	0	3 (1.1)
CS8	276 (98.6)	0	4 (1.4)
CS16	276 (98.6)	0	4 (1.4)
overall	2207 (98.5)	14 (0.6)	19 (0.9)

Table 5: Frequencies and percentages (in parentheses) of minimally specified, underspecified and overspecified descriptions in each condition in Experiment 2.

	2	4	8	16	Overall
C	1785 (415)	1737 (374)	1759 (382)	1740 (407)	1755 (394)
CS	1862 (491)	1887 (495)	1926 (501)	1953 (475)	1907 (491)
Overall	1824 (454)	1812 (445)	1842 (452)	1845 (454)	–

Table 6: Mean speech onset times (in milliseconds) and standard deviations (in parentheses) for minimimally specified descriptions, in each condition in Experiment 2.

Results

Table 6 displays mean speech onset times and standard deviations as a function of condition, as well as across conditions, for minimally specified descriptions. In contrast to Experiment 1, there is little prima facie evidence that speech onset time increases with the number of distractors overall. Closer inspection reveals the expected trend in the CS condition, where the mean onset time increases, especially for domains with 8 or more distractors. By contrast, the means for the C condition show a drop from the two-distractor to the four-distractor condition, with smaller fluctuations for distractor set sizes above four. This is made more explicit in Figure 6, which also indicates an interaction between Distractors and Properties.

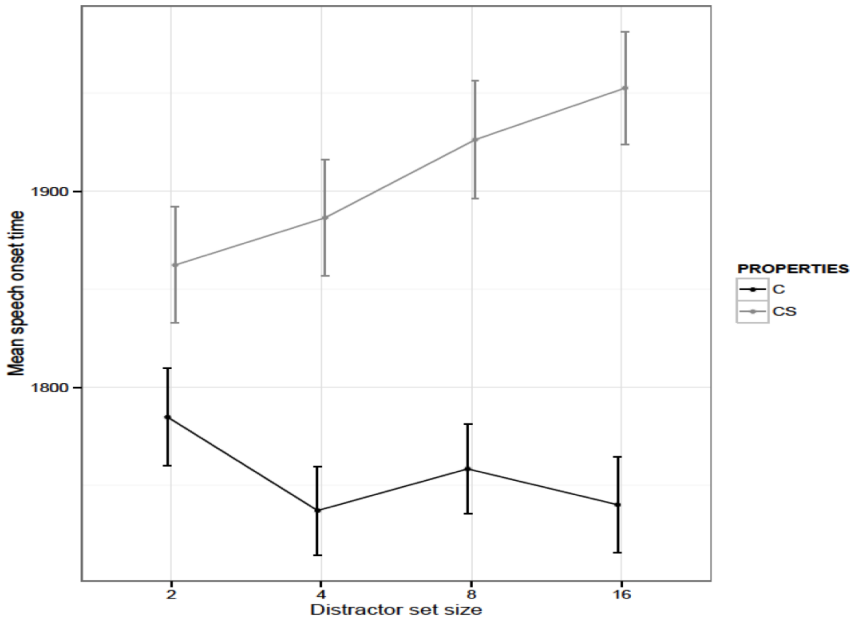


Figure 6. Mean onset times by number of distractors for minimally specified descriptions in the S and CS conditions. Error bars represent ±1 standard error of the mean

We turn now to the Linear Mixed Effects analyses, where we follow the same

strategy as per Experiment 1 in incrementally comparing models. All models converged with a maximal random effects structure. Once again, the fixed effect of Distractors is modelled numerically; both Properties and Distractors were centred. Possible item frequency effects are accounted for by comparing the best-fitting model to the same model with random intercepts and slopes for noun logarithmic frequency, calculated as per Experiment 1. Model comparisons are summarised in Table 7.

	Fixed effects	BIC	**Model χ^2**
0	Intercept only (baseline)	31651	–
1	Properties	31632	27.11^{**} (relative to model 0)
2	Distractors	31659	$0.06, ns$ (relative to model 0)
3	Properties + Distractors	31639	27.71^{**} (relative to model 0)
4	Properties × Distractors	31641	5.91^{*} (relative to model 3)

Table 7: Model goodness of fit statistics: (**) indicates that model has significantly better goodness of fit than the model indicated in parentheses at $p < .001$; (*) indicates better fit at $p < .05$.

The difference between the two levels of Properties (model 1) contributes significantly to explaining the variation in speech onset times, but distractor set size (model 2) does not: this model does not fit the data any better than the baseline. The addition of both fixed effects in model 3 outperforms the baseline model. Although this model fits the data better than model 2, which incorporates only the fixed effect of Distractors

($\chi^2 = 27.65, p < .001$), it is no better than the model containing only Properties (model 1; $\chi^2 < 1, ns$), suggesting that it is Properties that is playing the most important role in explaining the variance. Model 4, which incorporates the interaction, fits the data best, achieving a better fit than model 3. A comparison of this model to the same model incorporating random intercepts and slopes for log-transformed noun frequency revealed no significant difference in goodness of fit (BIC for model incorporating frequency: 32018; $\chi^2 = 27.4, p > 0.9$). Once again, we conclude that frequency did not exert an impact on speech onset times.

Parameter	**Estimate**	**Standard Err.**	t-value
Intercept	1840.10	37.43	49.15**
Properties	153.81	26.11	5.89**
Distractors	13.33	9.80	$1.36, ns$
Properties × Distractors	48.17	19.29	2.50*

Table 8: Details of the best-fitting model for the data in Experiment 2, incorporating fixed effects of Properties and Distractors and their interaction. (**) indicates significance at $p < .001$; (*) at $p < .05$.

The full details of model 4 are displayed in Table 8. Overall, an increase in the number of distractors does not result in an increase in speech onset time. However, the presence of a significant interaction shows that larger distractor set sizes had an impact in the CS condition, but not in the C condition. To investigate this further, we

carried out separate Linear Mixed Effects analyses on the C and CS data, using the same model-comparison strategy as before to compare models with Distractors as the sole fixed effect to a baseline model. The models are summarised in Table 9. All models included random intercepts and random slopes by participants and items.

Condition	**Fixed effects**	BIC	**Model χ^2**
C	Intercept only (baseline)	15594	–
	Distractors	15600	1.24, *ns*
CS	Intercept only (baseline)	15948	–
	Distractors	15942	6.03*

Table 9: Separate model goodness of fit statistics for the C and CS conditions: (*) indicates significance at $p < .05$;

As the table shows, a model that incorporates Distractors does not improve fit over the baseline in the C condition. Significance testing shows no main effect of Distractors ($t = 1.12, p > .2$). By contrast, the fixed effect of Distractors contributes to a signficantly better fit in the CS condition, with a main effect of Distractors ($t = 2.53, p = .01$). Thus, our results suggest no impact of domain size on speech onset time when a target referent is identified by a single property, in this case a highly contrastive difference in colour; on the other hand, domain size has a significant impact in case the target referent requires two properties in an identifying description.

Discussion

The results of Experiment 2 show that there is little effect of number of distractors on the time taken to initiate a referring expression, in case the property required is colour. By contrast, content determination involving a conjunction of properties (colour and size) is significantly slower, and an increase in the number of distractors in the domain results in an increase in speech onset time. By contrast, in Experiment 1, the impact of distractors was evident even in case a target referent could be identified using only its size.

There is a further notable difference between speech onset times for Experiment 2 and Experiment 1 in the CS condition, which was identical in both experiments: in Experiment 1, the mean latency over all levels of Distractors was 2073.27, compared to a mean of 1906.76 in Experiment 2 (see Tables 2 and 6). This difference is likely due to an effect of experimental context: peformance may have been influenced by the alternation of the CS condition with size in Experiment 1, and with colour in Experiment 2.

We explored this possibility by combining the data from Experiments 1 and 2. We compared the two models summarised in Table 10, the first of which included Distractors, Properties and their interaction, while the second also included a fixed effect of Item Order. Item Order was incorporated as a continuous predictor, with a number indicating the trial where an item was encountered in the course of the experiment. This was possible, since item orders were determined in advance, using an identical pseudo-random order in both experiments, so that every item occurred in the same order in the course of the experiment, irrespective of condition. A fixed effect of Experiment was also included

in both models as a between-groups factor; both models included random intercepts and slopes for Distractors, Properties and Order by participants (nested within Experiment) and items.

	Fixed effects	BIC	**Model χ^2**
1	Properties × Distractors × Experiment	56754	
2	Properties × Distractors × Experiment × Order	56771	16.40*

Table 10: Model goodness of fit statistics reflecting the impact of item order in the data from both experiments: (*) indicates that model has significantly better goodness of fit than the model indicated in parentheses at $p < .01$.

While the second model had a slightly higher BIC, likely due to the greater number of parameters compared to the first, it was significantly better at predicting the data than Model 1. Significance tests over Model 2 showed that there was a main effect of Experiment ($SE = 54.01; t = 2.60, p < .01$). This is unsurprising, given the observed difference in mean speech onset time between the two experiments. Table 11 summarises the remaining components of the model.

When the data from both experiments are combined, the main effects of Properties and Distractors approach, but do not reach significance, while Item Order exerts a significant main effect. Crucially, we observe a significant interaction between Properties and Distractors, confirming the observations, from the separate analyses of the two experiments, that the impact of domain size depends on the property combination

Parameter	**Estimate**	**Standard Err.**	t-value
Intercept	1905.65	32.45	58.72^{**}
Properties	99.53	21.04	$4.73^{\dagger}$
Distractors	31.58	7.72	$4.08^{\dagger}$
Item Order	-83.96	23.51	3.57^{**}
Properties × Distractors	24.90	7.48	3.33^{**}
Distractors × Item Order	-7.02	6.50	$1.08, ns.$
Properties × Item Order	-26.171	7.99	3.27^{**}
Distractors × Properties × Item Order	-7.357	7.38	$0.99, ns.$

Table 11: Details of the best-fitting model for the combined data in Experiments 1 and 2, incorporating fixed effects of Properties and Distractors and their interaction. (†) indicates that a fixed effect approaches, but does not reach, significance at $p < .05$; (**) indicates significance at $p < .001$; (*) at $p < .05$.

required for identification. An interaction of Item Order with Properties lends support to the conclusion that the gain in speed made by participants over a series of trials was dependent on which property was required to distinguish the target referent. This is further supported by studying the role of Item Order in each experiment individually. Individual models were obtained by adding the fixed effect of Item Order to the best-fitting model found for each experiment (model 3 in Table 3 and model 4 in Table 7). In Experiment 1, the model estimate for the main effect of order was $-109.61ms$, while it was -61.40 for Experiment 2. Both indicate a gain in speed, but this gain is higher in the

first Experiment, suggesting that here latencies were initially slower as speakers needed to scrutinise size differences, compared to Experiment 2, where the single-property condition involved an easily discriminable colour difference. The upshot was a greater gain in speed over the course of Experiment 1.

To summarise the findings, the current experiment showed that the effect of distractor set size on speech onset latencies depended on the properties used to distinguish a referent; in particular, we do not find a main effect of Distractors, but we do find an interaction with Properties, in contrast to the findings in Experiment 1, where size, rather than colour, was compared to the CS condition. Combining the data from both experiments confirms that the distractor set size effect is dependent on the properties required to plan a description; furthermore, as participants become more efficient at the task, as reflected by the impact of item order, the extent to which there is a gain also depends on the properties required for an identifying description.

General discussion

Our starting point in this paper was the observation that conceptualisation during reference production can be modelled as a search process, a view adopted by computational models developed within the field of Referring Expression Generation (REG; Krahmer & van Deemter, 2012). Furthermore, some insights from these models and from empirical work on reference production converge with insights from several decades of research on visual search, especially where this work has shed light on the efficiency of search processes, from so-called 'pop-out' search at one extreme, through various de-

grees of difficulty arising from the conditions present in the visual stimulus. At the same time, there is a difference in emphasis between the two bodies of work. Search during reference production begins from a known target referent and proceeds through a search space defined by the target's properties and their combinations. In visual search, the starting point is usually a template, or a description of a target, and the aim is to verify its presence or absence.

The primary aim of this paper was to test the predictions of REG models, which predict that the efficiency of search for properties of a target referent will be affected both by the number of properties required to identify it, and by the number of distractors to which the target needs to be compared. The results only lend partial support to these predictions. As REG models predict, increasing distractor set size makes content determination less efficient in case a referent needs to be distinguished by a conjunction of colour and size. This is reminiscent of early results in visual search (e.g. Treisman & Gelade, 1980) showing that conjunction search exhibits a linear dependency on the number of distractors. However, contrary to the predictions of models, the effect of increasing distractor set size where a referent can be distinguished on the basis of a single property depends on the property under consideration. Where the property in question was highly discriminatory (colour in Experiment 2), search latencies were unaffected by distractor set size (reminiscent of a 'pop-out' effect in some visual search studies). With a less discriminatory property – such as the size contrast in Experiment 1 – search is affected by domain size. This is further supported by the observation that participants made gains in efficiency in the course of the experiment, but the extent of these gains

differed among these two cases.

In the remainder of this section, we outline some of the implications of these findings for our understanding of reference production, first by holding up the computational models we have considered against our experimental results, as well as results from recent work on visual search, and then by considering the prospects for models that go beyond artificial visual domains.

Implications for models of reference production

Two of the assumptions of the models we have considered deserve further scrutiny in light of the results of our experiments.

The first assumption is that conceptualisation or content determination is performed against a knowledge base where entities and their properties are represented. This allows these models to assume a separation between the speaker's initial identification of a target's properties on the one hand – for example, her knowledge of the target's colour, shape, or size – and the search through those properties on the other. It is only the latter that is considered central to conceptualisation or content determination. This assumption is clear in the formulation of the algorithm schema outlined in Figure 2.

There are approaches which do pay closer attention to the establishment of the initial set of properties on the basis of which search will be carried out, especially where these properties can be numerically represented (as in the case of height or width of an object; van Deemter, 2006) or where these properties consist of spatial landmarks that can be used to identify a target in a relation (such as *to the right of X*; cf. J. D. Kelleher

& Kruijff, 2006; Elsner, Rohde, & Clarke, 2014; Clarke, Elsner, & Rohde, 2015). In the latter cases, the salience of a landmark plays an explicit role in determining whether it is included in a description.

Common to these approaches is the notion that visual salience can be used to prioritise information during search, a view that harks back to an early body of work on language generation (e.g. Arbib, Conklin, & Hill, 1987; Novak, 1987, *inter alia*). What is missing from this picture is an account of how salience itself can inform that part of the content determination process whereby a property is determined to be relevant to a distinguishing description. This leads us to the second assumption underlying the models under discussion: while properties can be prioritised during search on the basis of various heuristics, including discriminatory power (Dale, 1989; Frank & Goodman, 2012), salience or preference (Dale & Reiter, 1995; Gatt et al., 2011; Mitchell et al., 2013), the inclusion of a property in a description ultimately depends on a comparison with the distractors in the relevant domain, so that a property is included if it is found to be contrastive. This holds irrespective of the property under consideration. Such a Gricean view of content determination has continued to dominate REG models, though once again, work which pays closer attention to the interface between vision and language production (e.g. Kazemzadeh et al., 2014) has tended to weaken this Gricean orientation. Thus, as far as contrastiveness is concerned, 'all properties are equal', in the sense that they are subject to the same treatment, even if they might be considered first.

To a first approximation, one could maintain this process model, with one alteration, and still account for the experimental results reported here. Based on the finding

that where colour is the property required to distinguish an intended referent (Experiment 2), there is little evidence of comparison between target and distractors, the model could be altered to first select colour, then check whether the resulting description is distinguishing, proceeding to search for other discriminatory properties should this not be the case. This account would remain faithful to the notion of a preference order, in which colour takes precedence over size or other prototypically gradable properties, also accounting for the redundant use of preferred properties in overspecified descriptions (Pechmann, 1989; Deutsch & Pechmann, 1982; Eikmeyer & Ahlsèn, 1996; Koolen et al., 2011; Belke, 2006; Arts, 2004). It would also appear to address findings in the vision literature that confirm the centrality of colour to object recognition (e.g. Wurm et al., 1993; Naor-Raz et al., 2003) and early visual processes (e.g., Itti & Koch, 2001; Wolfe & Horowitz, 2004; Wolfe, 2010, among others). Such a privileged treatment of colour is also a feature of some recent stochastic REG models (Mitchell et al., 2013; Gatt et al., 2011).

However, we suggest that this account wouldn't make an algorithm a psychologically realistic process model, for two reasons. First, the dichotomy between colour and other properties such as size has been qualified in recent years, based on evidence that the preference or salience of colour is dependent on its discriminability (Viethen et al., 2012), the homogeneity of the scene (Koolen, Goudbeek, & Krahmer, 2013), as well as colour typicality in relation to the type of object being described (Sedivy, 2003; Westerbeek et al., 2015). Similar evidence has been reported for size contrasts, which are in any case also subject to salience-based processing (Stuart et al., 1993; Treisman & Gormican,

1988; Busch & Müller, 2004). Where such contrasts are highly salient, size is no longer as dispreferred by speakers, suggesting that the contrast is easier to identify and encode during the content determination process (Hermann & Deustch, 1976; Levelt, 1989; van Gompel et al., 2014).

These findings converge with models of visual search proposed in the wake of the foundational work of Treisman and Gelade (1980): for example, Duncan and Humphreys (1989) proposed that similarity or contrastiveness is central to determining search efficiency. The Biased Competition model (Desimone & Duncan, 1995; Desimone, 1998) similarly assumes that attention is drawn to salient regions as a result of competition taking place across the visual field. Wolfe's Guided Search model (Wolfe, 1994, 2007) also emphasises a reliance on salient features to guide attentional deployment to salient regions.

Turning back to our experimental results, the difference found between the single-property cases – colour on the one hand, and size on the other – could be accounted for on the basis that the colour contrasts used in our experiment were relatively stark and easy to perceive, compared to the size contrasts. This would account for the absence of an effect of distractor set size with colour-only descriptions. This more nuanced view, we argue, should underlie future models of conceptualisation in reference production, where salience and discriminability not only determine the *order* in which properties of a target referent are considered for inclusion in a description, but also inform the process of selection itself. A highly discriminable feature or set of features would serve to draw attention to a specific region in a visual scene, thereby circumscribing the relevant

portion of the scene within which a target needs to be compared to its distractors. The extent of this comparison would depend on whether the feature uniquely characterises the target or not. Some eye-tracking studies lend preliminary support to this account. For example, Brown-Schmidt and Tanenhaus (2006) found that the use of a size adjective for a target referent is more likely in the presence of a size contrast, following a fixation to a distractor of differing size. A more general finding is that during visual search, there are more saccades to regions of a scene containing objects which are visually similar to a target, both in complex, real-world scenes (Hwang et al., 2011) and in artificial displays of realistic objects (Alexander & Zelinsky, 2011), although these findings also need to be discussed in light of other findings concerning semantic similarity and global scene properties (see below).

Establishing the validity of a salience-based account of the sort sketched here requires much more research that explicitly manipulates the degree of salience of different features, using methodologies, such as eye-tracking, which shed a direct light on the process of domain circumscription during reference production.

What this tentative account does not address is the differences, such as they are, between single property and conjunctive descriptions, where our experiments suggest speakers are slower and may be engaging in serial comparison to the distractors. However, the findings of our experiments do suggest that the single-property versus conjunction distinction is also not as crisp as assumed by classical models of visual search such as FIT (Treisman & Gelade, 1980), or models of reference production compatible with the schema in Figure 2. Distractor set size impacted production latencies in the

conjunctive case in both experiments, but made search less efficient only for size in the single-property case. When the results from both experiments are combined, the main effect of Distractors only approaches significance, while its interaction with Properties exerts a highly significant effect on speech onset times. A plausible interpretation of these patterns is that where a salient colour contrast existed, it was discerned relatively quickly. In case colour alone was sufficient for identification, search could terminate; in case an additional property was required – in this case, size, which we have suggested was less salient in our manipulation – some comparison to the distractors was needed to determine the value for the target referent. Under this account, in the CS conditions of both experiments, the colour contrast could have supported a subset search, where the target needed to be compared to distractors on the basis of size, but only within the subset of distractors that had the same colour as the target. This view finds some support in the visual search literature, which has shown that the effiency of conjunctive search is enhanced in visual displays that afford subset search strategies or, more generally, the formation of perceptual groups (e.g. Nakayama & Silverman, 1986; He & Nakayama, 1995; Nakayama & Joseph, 1998; Friedman-Hill & Wolfe, 1995; Nordfang & Wolfe, 2014).

Conclusion: Beyond static, artificial scenes?

The experiments in this paper were designed to address questions arising from reference production data and models. As we noted at the outset, much of this work has relied on visual displays which are artificially constructed and the present paper was

no exception. This is a property shared with many visual search experiments. While it has the obvious benefit of enabling researchers to control the relevant conditions in the display, it has also led to models which focus exclusively on the relationship between target and distractors, usually on the basis of exclusively visual properties, in domains where speakers' knowledge and expectations can be ignored (but see Stoia & Shockley, 2006; Garoufi & Koller, 2013; Elsner et al., 2014, for examples of models that deviate somewhat from this norm). While this general picture is changing, as the problem of automatically generating descriptions of real-world scenes receives more attention (e.g. Farhadi et al., 2010; Feng & Lapata, 2010; Yang, Teo, Daume, & Aloimonos, 2011; Mitchell et al., 2012; Elliott & Keller, 2013; Yatskar et al., 2014; Kulkarni et al., 2013, among many others), there is as yet very little work that specifically addresses reference production in such scenes from a computational perspective, or only does so by focussing attention on low-level features. Thus, Kazemzadeh et al. (2014) propose a model for referring expression generation in complex photographs which exploits visual features, but does not incorporate knowledge of other factors that come into play when we view such scenes. In this concluding section, we speculate on the challenges that arise for models of reference production when such factors are explicitly considered, based on recent research in visual search.

Evidence that search patterns are not exclusively guided by visual properties of a scene has been forthcoming from experiments showing both that visual search can be made less efficient by semantically related distractors (e.g. Belke et al., 2008) and that semantic relationships between objects in a real-world scene (measured, for example,

by distributional semantic models such as Latent Semantic Analysis Landauer, Foltz, & Laham, 1998) are better predictors of eye-movement patterns than relationships based on visual similarity alone, though the latter also have some predictive power (Hwang et al., 2011). In a related vein, eye-movement research by Henderson et al. (2007) showed that the fixated regions of a complex scene tended to be those with lower intensity and high local contrast; however, these were also regions that were independently judged as having high semantic content. The role of saliency in vision has also been shown to depend on the nature of the task. For instance, Einhäuser and Koch (2008) found that the degree to which observers' saccades were determined by visually salient regions differed between a free viewing task, a template search and an odd-one-out detection task. Similar conclusions have been reached on the basis of studies showing strong selection biases that are not due to bottom-up factors, but to expected reward based on previous experience and to previous selection history (see Awh et al., 2012, and references therein).

Thus, low-level properties of a scene, including the salience of regions based on their features, interact with processes of attention and selection in more complex ways than envisaged by models based on visual search in artificial displays. Rather, they highlight the role of global and contextual effects at multiple levels, including task and semantic relationships. For example, in searching for a particular object in a scene, an observer's attention is likely to be guided not only by the task itself, but by their knowledge of the structure of such scenes, based among other things on expected regularities (Oliva & Torralba, 2007). For instance, the Contextual Guidance Model (Torralba et

al., 2006) addresses this by modelling the likelihood that a target is located in a region of a scene in a Bayesian framework, as a function both of locally salient features and global scene priors learned from past experience. A comparison of the predictions of the model to eye-movement data has shown that the inclusion of contextual priors improves model accuracy, compared to a model that only incorporates salience. Interestingly, excluding local information and including only contextual priors in the model shows a much smaller decrease in accuracy, suggesting that in real-world scenes, it is context that plays the dominant role in guiding attention. A different model, SUN (Kanan et al., 2009) highlights the role of prior knowledge of the target class and appearance of the target. Under this model, the likelihood that a target is present at some point in the visual field is contingent on bottom-up saliency, target appearance and target location. This model has also outperformed a purely bottom-up model in predicting fixations.

How might these findings alter our view of reference production? Incorporating contextual and object-based knowledge would need to take into account the differences between visual search tasks and reference production. The question for reference production raised by models such as that of Torralba et al. (2006) or Kanan et al. (2009) is how contextual or object-based priors can influence what might be said about a referent, rather than how quickly the entity might be detected.

Consider a common scenario, such as an office scene. In order to refer to an object in the scene, such as a pen or a telephone, a speaker's conceptualisation of the referent is likely to be informed by such factors as its typical location (on the desk), whether it would be expected to be the only such object in the scene (an office may contain several

pens, but might be less likely to have more than one telephone), as well as deviations from such expectations (as in the case where the telephone is on the floor). What is selected, as well as the amount of information conveyed in a description, would be expected to change as a function of such deviations (the findings cited above, to the effect that the likelihood with which 'preferred' properties are used changes as a function of their predictability with respect to a referent, are compatible with this view; cf. Sedivy, 2003; Westerbeek et al., 2015). Thus, a speaker might choose to refer to *the red pen* if its colour were salient among similar objects in the relevant portion of the scene. On the other hand, a deviation from the usual location of the target referent might alter the referential strategy altogether (yielding, for example, *the pen on the chair*).

In the previous section, we informally sketched an alternative to current REG models, based on a graded salience mechanism underlying property selection and comparison during content determination. Consideration of contextual factors opens up an avenue for research into how such a salience-based mechanism is modulated by prior knowledge and expectations, and how this impacts planning and choice.

An explicit account of the role of contextual and world knowledge remains elusive in reference production models (but see Kutlak, van Deemter, & Mellish, 2012, for a computational account of communal common ground relying on general knowledge). Just as experimental and modelling work using classical visual search paradigms has yielded interesting convergences with reference production, more recent work incorporating contextual priors offers insights that can bring reference production models closer to real-world language production tasks. A greater synergy between research

on language production and research on vision can enhance our understanding of how speakers conceptualise referents in visual scenes.

References

Achanta, R., Hemamiz, S., Estraday, F., & Süsstrunky, S. (2009). Frequency-tuned salient region detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'09)* (pp. 1597–1604). doi: 10.1109/CVPRW.2009.5206596

Alexander, R. R. G., & Zelinsky, G. J. G. (2011). Visual similarity effects in categorical search. *Journal of vision*, *11*(8), 1–15. doi: 10.1167/11.8.9.Introduction

Altmann, G. T., & Kamide, Y. (1999, dec). Incremental interpretation at verbs: restricting the domain of subsequent reference. *Cognition*, *73*(3), 247–64.

Appelt, D. (1985). Planning English referring expressions. *Artificial Intelligence*, *26*(1), 1–33.

Arbib, M. A., Conklin, E. J., & Hill, J. C. (1987). *From Schema Theory to Language*. Oxford: Oxford University Press.

Arts, A. (2004). *Overspecification in Instructive Texts* (Unpublished doctoral dissertation). Tilburg University.

Awh, E., Belopolsky, A. V., & Theeuwes, J. (2012). Top-down versus bottom-up attentional control: A failed theoretical dichotomy. *Trends in Cognitive Sciences*, *16*(8), 437–443. doi: 10.1016/j.tics.2012.06.010

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

Bates, D., Maechler, M., & Bolke, B. (2014). *lme4: Linear mixed-effects models using S4 classes.*

Bauer, B., Jolicoeur, P., & Cowan, W. B. (1996). Visual search for colour targets that are or are not linearly separable from distractors. *Vision Research*, *36*(10), 1439–1465. doi:

10.1016/0042-6989(95)00207-3

Belke, E. (2006, jul). Visual determinants of preferred adjective order. *Visual Cognition*, *14*(3), 261–294. doi: 10.1080/13506280500260484

Belke, E., Humphreys, G. W., Watson, D. G., Meyer, A. S., & Telling, A. L. (2008). Top-down effects of semantic knowledge in visual search are modulated by cognitive but not perceptual load. *Perception and Psychophysics*, *70*(8), 1444–1458. doi: 10.3758/PP

Belke, E., & Meyer, A. S. (2002, apr). Tracking the time course of multidimensional stimulus discrimination: Analyses of viewing patterns and processing times during "same"-"different" decisions. *European Journal of Cognitive Psychology*, *14*(2), 237–266. doi: 10.1080/09541440143000050

Bohnet, B., & Dale, R. (2005). Viewing referring expression generation as search. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'05)* (pp. 1004–1010).

Brown-Schmidt, S., & Tanenhaus, M. K. (2006, may). Watching the eyes when talking about size: An investigation of message formulation and utterance planning. *Journal of Memory and Language*, *54*(4), 592–609. doi: 10.1016/j.jml.2005.12.008

Busch, A., & Müller, H. J. (2004). The Ebbinghaus illusion modulates visual search for size-defined targets: evidence for preattentive processing of apparent object size. *Perception & psychophysics*, *66*(3), 475–495. doi: 10.3758/BF03194895

Campana, E., Tanenhaus, M. K., Allen, J. F., & Remington, R. (2010, sep). Natural discourse reference generation reduces cognitive load in spoken systems. *Natural Language Engineering*, *17*(03), 311–329. doi: 10.1017/S1351324910000227

Chambers, C. (2002, jul). Circumscribing Referential Domains during Real-Time Language Comprehension. *Journal of Memory and Language*, *47*(1), 30–49. doi: 10.1006/jmla.2001.2832

Chiu, E. M., & Spivey, M. J. (2012). The Role of Preview and Incremental Delivery on Vi-

sual Search. In *Proceedings of the 34th Annual Meeting of the Cognitive Science Society (CogSci'12)* (pp. 216–221). Austin, TX: Cognitive Science Society.

Clarke, A. D. F., Elsner, M., & Rohde, H. (2015). Giving good directions: order of mention reflects visual salience. *Frontiers in Psychology*, *6*, 1793. doi: 10.3389/fpsyg.2015.01793

Dale, R. (1989). Cooking up referring expressions. In *Proceedings of the 27th annual meeting on Association for Computational Linguistics (ACL'89)* (pp. 68–75). Vancouver, BC: Association for Computational Linguistics. doi: 10.3115/981623.981632

Dale, R., & Reiter, E. (1995, apr). Computational Interpretations of the Gricean Maxims in the Generation of Referring Expressions. *Cognitive Science*, *19*(2), 233–263. doi: 10.1207/s15516709cog1902_3

Desimone, R. (1998). Visual attention mediated by biased competition in extrastriate visual cortex. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *353*(1373), 1245–55. doi: 10.1098/rstb.1998.0280

Desimone, R., & Duncan, J. S. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, *18*(1), 193–222. doi: 10.1146/annurev.ne.18.030195.001205

Deutsch, W., & Pechmann, T. (1982). Social interaction and the development of definite descriptions. *Cognition*, *11*, 159–184.

Duncan, J., & Humphreys, G. W. (1989). Visual search and stimulus similarity. *Psychological Review*, *96*(3), 433–458.

Eikmeyer, H. J., & Ahlsèn, E. (1996). The cognitive process of referring to an object: A comparative study of {G}erman and {S}wedish. In *Proceedings of the 16th Scandinavian Conference on Linguistics.*

Einhäuser, W., & Koch, C. (2008). Task-demands can immediately reverse the effects of sensory-driven saliency in complex visual stimuli. *Journal of Vision*, *8*(2), 1–19. doi: 10.1167/8.2.2.Introduction

Elliott, D., & Keller, F. (2013). Image Description using Visual Dependency Representations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP'13)* (pp. 1292–1302).

Elsner, M., Rohde, H., & Clarke, A. D. F. (2014). Information Structure Prediction for Visual-world Referring Expressions. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL'14)* (pp. 520–529). Gothenburg, Sweden: Association for Computational Linguistics.

Engelhardt, P., Bailey, K., & Ferreira, F. (2006, may). Do speakers and listeners observe the Gricean Maxim of Quantity? *Journal of Memory and Language*, *54*(4), 554–573. doi: 10.1016/j.jml.2005.12.009

Erdem, E., & Erdem, A. (2013). Visual saliency estimation by nonlinearly integrating features using region covariances. *Journal of vision*, *13*(4), 11, 1–20. doi: 10.1167/13.4.11

Farhadi, A., Hejrati, M., Sadeghi, M. A., Young, P., Rashtchian, C., Hockenmaier, J., & Forsyth, D. (2010). Every picture tells a story: Generating sentences from images. In *Proceedings of the 11th European Conference on Computer Vision (ECCV'10)* (Vol. 6314 LNCS, pp. 15–29). Berlin and Heidelberg: Springer. doi: 10.1007/978-3-642-15561-1_2

Feng, Y., & Lapata, M. (2010). How many words is a picture worth? Automatic caption generation for news images. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL'10)* (pp. 1239–1249). Uppsala, Sweden: Association for Computational Linguistics.

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

Found, A. (1998). Parallel coding of conjunctions in visual search. *Perception {&} psychophysics*, *60*(7), 1117–1127.

Frank, M. C., & Goodman, N. D. (2012, may). Predicting pragmatic reasoning in language

games. *Science (New York, N.Y.)*, *336*(6084), 998. doi: 10.1126/science.1218633

Friedman-Hill, S., & Wolfe, J. M. (1995). Second-order parallel processing: visual search for the odd item in a subset. *Journal of Experimental Psychology: Human Perception and Performance*, *21*(3), 531–551. doi: 10.1037/0096-1523.21.3.531

Garoufi, K., & Koller, A. (2013). Generation of effective referring expressions in situated context. *Language and Cognitive Processes*, *29*(8), 986-1001. doi: 10.1080/01690965.2013.847190

Gatt, A., van Gompel, R. P. G., 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 psycholinguistic approaches to reference (PRE-CogSci'11).*

Gibson, B. S., Eberhard, K. M., & Bryant, T. A. (2005). Linguistically mediated visual search: the critical role of speech rate. *Psychonomic Bulletin & Review*, *12*(2), 276–281.

Griffin, Z. M., & Bock, K. (2000). What the eyes say about speaking. *Psychological science*, *11*, 274–279. doi: 10.1111/1467-9280.00255

Haslam, N., Porter, M., & Rothschild, L. (2001). Visual search: efficiency continuum or distinct processes? *Psychonomic bulletin & review*, *8*(4), 742–746. doi: 10.3758/BF03196212

He, Z. J., & Nakayama, K. (1995). Surface vs Features in visual search. *Nature*, *359*, 231–233.

Henderson, J. M., Brockmole, J. R., Castelhano, M. S., & Mack, M. (2007). Visual Saliency Does Not Account for Eye Movements During Visual Search in Real World Scenes. In R. van Gompel, M. Fischer, W. Murray, & H. R.L. (Eds.), *Eye Movements: A Window on Mind and Brain* (pp. 1–41). London: Elsevier. doi: 10.1167/9.3.6.

Hermann, T., & Deustch, W. (1976). *Psychologie der Objektbenennung.* Bern: Huber Verlag.

Hwang, A. D., Wang, H.-C., & Pomplun, M. (2011). Semantic guidance of eye movements in real-world scenes. *Vision Research*, *51*(10), 1192–1205. doi: 10.1016/j.visres.2011.03.010

Itti, L. (2005). Models of bottom-up attention and saliency. In L. Itti, G. Rees, & J. K. Tsotsos

(Eds.), *Neurobiology of Attention* (pp. 576–582). San Diego, Ca.: Elsevier.

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience*, *2*(3), 194–203. doi: 10.1038/35058500

Jordan, P. W., & Walker, M. a. (2005). Learning content selection rules for generating object descriptions in dialogue. *Journal of Artificial Intelligence Research*, *24*, 157–194. doi: 10.1613/jair.1591

Kanan, C., Tong, M. H., Zhang, L., & Cottrell, F. W. (2009). SUN: Top-down saliency using natural statistics. *Visual Cognition*, *17*(6-7), 979–1003. doi: 10.1016/j.micinf.2011.07.011.Innate

Kazemzadeh, S., Ordonez, V., Matten, M., & Berg, T. L. (2014). ReferItGame: Referring to Objects in Photographs of Natural Scenes. In *Proceedings of the 2014 Conrerence on Empirical Methods in Natural Language Processing (EMNLP'14)* (pp. 787–798). Doha, Qatar: Association for Computational Linguistics.

Kelleher, J., Costello, F., & Van Genabith, J. (2005). Dynamically structuring, updating and interrelating representations of visual and linguistic discourse context. *Artificial Intelligence*, *167*, 62–102. doi: 10.1016/j.artint.2005.04.008

Kelleher, J. D., & Kruijff, G.-J. 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 (COLING-ACL'06)* (pp. 1041–1048). Sydney, Australia: Association for Computational Linguistics.

Knoeferle, P., Crocker, M. W., Scheepers, C., & Pickering, M. J. (2005, feb). The influence of the immediate visual context on incremental thematic role-assignment: evidence from eye-movements in depicted events. *Cognition*, *95*(1), 95–127. doi: 10.1016/j.cognition.2004.03.002

Koolen, R., Gatt, A., Goudbeek, M., & Krahmer, E. (2011). Factors causing overspecification in definite descriptions. *Journal of Pragmatics*, *43*, 3231–3250.

Koolen, R., Goudbeek, M., & Krahmer, E. (2013). The Effect of Scene Variation on the Redundant Use of Color in Definite Reference. *Cognitive Science*, *37*, 395–411.

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

Kulkarni, G., Premraj, V., Ordonez, V., Dhar, S., Li, S., Choi, Y., ... Berg, T. L. (2013). Baby talk: Understanding and generating simple image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(12), 2891–2903. doi: 10.1109/TPAMI.2012.162

Kutlak, R., van Deemter, K., & Mellish, C. (2012). Corpus-based metrics for assessing communal common ground. In *Proceedings of the 34th Annual Meeting of the Cognitive Science Society (CogSci'12)* (pp. 1834–1839). Austin, TX: Cognitive Science Society.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2014). *lmerTest: Tests for random and fixed effects for linear mixed effect models.*

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis,. *Discourse Processes*, *25*, 259–284. doi: 10.3758/BRM.41.3.944

Levelt, W. M. (1989). *Speaking: From Intention to Articulation.* Cambridge, MA: MIT Press.

Levelt, W. M. (1999). Producing spoken language: a blueprint of the speaker. In C. Brown & P. Hagoort (Eds.), *The Neurocognition of Language* (pp. 83–122). Oxford and London: Oxford University Press.

Lewandowsky, S., & Farrell, S. (2010). *Computational Modeling in Cognition: Principles and Practice.* London: Sage Publications Inc.

Malcolm, G. L., & Henderson, J. M. (2009). The effects of target template specificity on visual search in real-world scenes: Evidence from eye movements. *Journal of Vision*, *9*(2009),

1–13. doi: 10.1167/9.11.8.Introduction

Mitchell, M., Dodge, J., Goyal, A., Yamaguchi, K., Stratos, K., Han, X., ... Daume III, H. (2012). Midge: Generating Image Descriptions From Computer Vision Detections. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL'12)* (pp. 747–756). Avignon, France: Association for Computational Linguistics.

Mitchell, M., van Deemter, K., & Reiter, E. (2013). Generating Expressions that Refer to Visible Objects. In *Proceedings of the meeting of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL'13)* (pp. 1174–1184). Atlanta, Georgia: Association for Computational Linguistics.

Nakayama, K., & Joseph, J. S. (1998). Attention, pattern recognition and pop-out in visual search. In *The Attentive Brain* (pp. 279–298).

Nakayama, K., & Silverman, G. H. (1986). Serial and parallel processing of visual feature conjunctions. *Nature*, *320*(6059), 264–265. doi: 10.1038/320264a0

Naor-Raz, G., Tarr, M. J., & Kersten, D. (2003). Is color an intrinsic property of object representation? *Perception*, *32*(6), 667–680. doi: 10.1068/p5050

Neisser, U. (1967). *Cognitive Psychology.* New York: Appleton-Century-Crofts.

Nordfang, M., & Wolfe, J. M. (2014). Guided search for triple conjunctions. *Attention, perception & psychophysics*, *76*(6), 1535–59. doi: 10.3758/s13414-014-0715-2

Novak, H.-J. (1987). Strategies for generating coherent descriptions of object movements in street scenes. In *Natural Language Generation: New Results in Artificial Intelligence, Psychology and Linguistics.* Dordrecht: Nijhoff & NATO Scientific Affairs Division.

Olds, E. S., Cowan, W. B., & Jolicoeur, P. (2000a). Partial orientation pop-out helps difficult search for orientation. *Percept.Psychophys.*, *62*(7), 1341–1347.

Olds, E. S., Cowan, W. B., & Jolicoeur, P. (2000b). The time-course of pop-out search. *Vision*

Research, *40*, 891–912.

Olds, E. S., Cowan, W. B., & Jolicoeur, P. (2000c). Tracking visual search over space and time. *Psychonomic bulletin & review*, *7*(2), 292–300.

Olds, E. S., & Fockler, K. A. (2004). Does previewing one stimulus feature help conjunction search? *Perception*, *33*, 195–216. doi: 10.1068/p5162

Oliva, A., & Torralba, A. (2007, dec). The role of context in object recognition. *Trends in cognitive sciences*, *11*(12), 520–527. doi: 10.1016/j.tics.2007.09.009

Palmer, J. (1994). Set-size effects in visual search: the effect of attention is independent of the stimulus for simple tasks. *Vision Research*, *34*(13), 1703–1721. doi: 10.1016/0042-6989(94)90128-7

Palmer, J. (1995). Attention in visual search: Distinguishing four causes of a set-size effect. *Current Directions in Psychological Science*, *4*(4), 118–123.

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

Protopapas, A. (2007). Check Vocal: A program to facilitate checking the accuracy and response time of vocal responses from DMDX. *Behavior Research Methods*, *39*(4), 859–862.

Reali, F., Spivey, M. J., Tyler, M. J., & Terranova, J. (2006). Inefficient conjunction search made efficient by concurrent spoken delivery of target identity. *Perception & psychophysics*, *68*(6), 959–74. doi: 10.3758/BF03193358

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

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.

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

Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory*, *6*(2), 174–215.

Sobel, K. V., & Cave, K. R. (2002). Roles of salience and strategy in conjunction search. *Journal of Experimental Psychology: Human Perception and Performance*, *28*(5), 1055–1070. doi: 10.1037//0096-1523.28.5.1055

Spivey, M. J., Tyler, M. J., Eberhard, K. M., & Tanenhaus, M. K. (2001, jul). Linguistically mediated visual search. *Psychological science*, *12*(4), 282–6.

Stoia, L., & Shockley, D. (2006). Noun phrase generation for situated dialogs. In *Proceedings of the Fourth International Natural Language Generation Conference (INLG'06)* (pp. 81–88).

Stuart, G. W., Bossomaier, T. R. J., & Johnson, S. (1993). Preattentive processing of object size: implications for theories of size perception. *Perception*, *22*(10), 1175–1193. doi: 10.1068/p221175

Sun, R. (2008). *The Cambridge Handbook of Computational Psychology*. Cambridge: Cambridge University Press.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*(5217), 1632–1634.

Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review*, *113*(4), 766–786. doi: 10.1037/0033-295X.113.4.766

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

Treisman, A., & Gormican, S. (1988). Feature Analysis in Early Vision: Evidence From Search

Asymmetries. *Psychological Review*, *95*(1), 15–48.

van Deemter, K. (2006). Generating referring expressions that involve gradable properties. *Computational Linguistics*, *32*(2), 195–222. doi: 10.1162/coli.2006.32.2.195

van Deemter, K., Gatt, A., van Gompel, R. P. G., & Krahmer, E. (2012, apr). Towards a computational psycholinguistics of reference production. *Topics in cognitive science*, *4*(2), 166–83. doi: 10.1111/j.1756-8765.2012.01187.x

van Gompel, R. P., Gatt, A., Krahmer, E., & van Deemter, K. (2014). Overspecification in reference: Modelling size contrast effects. In *Proceedings of the 2014 conference on Architectures and Mechanisms in Language Processing (AMLAP'14).*

Vicente, K. L., & Wang, J. H. (1998). An ecological theory of expertise effects in memory recall. *Psychological Review*, *105 (1)*, 33–57.

Vickery, T. J., King, L.-W., & Jiang, Y. (2005). Setting up the target template in visual search. *Journal of vision*, *5*(1), 81–92. doi: 10.1167/5.1.8

Viethen, J., Goudbeek, M., & Krahmer, E. (2012). The impact of colour difference and colour codability on reference production. In *Proceedings of the 34th Annual Meeting of the Cognitive Science Society (CogSci'12)* (pp. 1084–1089). Austin, TX: Cognitive Science Society.

Walther, D., & Koch, C. (2006). Modeling attention to salient proto-objects. *Neural Networks*, *19*(9), 1395–1407. doi: 10.1016/j.neunet.2006.10.001

Westerbeek, H., Koolen, R., & Maes, A. (2015). Stored object knowledge and the production of referring expressions: the case of color typicality. *Frontiers in Psychology*, *6*(July), 1–12. doi: 10.3389/fpsyg.2015.00935

Wolfe, J. M. (1994). Guided Search 2 . 0 A revised model of visual search. *Psychonomic Bulletin and Review*, *1*(2), 202–238.

Wolfe, J. M. (1998). What Can 1 Million Trials Tell Us About Visual Search? *Psychological*

Science, *9*(1), 33–39. doi: 10.1111/1467-9280.00006

Wolfe, J. M. (2007). Guided Search 4.0 Current Progress With a Model of Visual Search. In W. D. Gray (Ed.), *Integrated Models of Cognitive Systems* (Vol. 1, pp. 99–119). New York: Oxford University Press. doi: 10.1167/1.3.349

Wolfe, J. M. (2010). Visual search. *Current Biology*, *20*(8), R346–9. doi: 10.1016/j.cub.2010.02.016

Wolfe, J. M., Cave, K. R., & Franzel, S. L. (1989). Guided search: an alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *15*(3), 419–433. doi: 2527952

Wolfe, J. M., & Horowitz, T. S. (2004). What attributes guide the deployment of visual attention and how do they do it? *Nature Reviews Neuroscience*, *5*(6), 495–501. doi: 10.1038/nrn1411

Wurm, L. H., Legge, G. E., Isenberg, L. M., & Luebker, a. (1993). Color improves object recognition in normal and low vision. *Journal of experimental psychology. Human perception and performance*, *19*(4), 899–911.

Yang, Y., Teo, C. L., Daume, H., & Aloimonos, Y. (2011). Corpus-Guided Sentence Generation of Natural Images. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP'11)* (pp. 444–454). Edinburgh, Scotland: Association for Computational Linguistics.

Yatskar, M., Galley, M., Vanderwende, L., & Zettlemoyer, L. (2014). See No Evil , Say No Evil : Description Generation from Densely Labeled Images. In *Proceedings of the Third Joint Conference on Lexical and Computation Semantics (*SEM).*