Visual complexity and its effects on referring expression generation

Micha Elsner[*] Alasdair Clarke[†] Hannah Rohde[‡]

April 24, 2017

Keywords: referring expression generation; psycholinguistics; sentence processing; visual search

[*]Department of Linguistics, The Ohio State University
[†]Department of Psychology, University of Essex
[‡]Department of Linguistics and English Language, University of Edinburgh

Abstract

Speakers' perception of a visual scene influences the language they use to describe it—which objects they choose to mention and how they characterize the relationships between them. We show that visual complexity can either delay or facilitate description generation, depending on how much disambiguating information is required and how useful the scene's complexity can be in providing, for example, helpful landmarks. To do so, we measure speech onset times, eye gaze, and utterance content in a reference production experiment in which the target object is either unique or non-unique in a visual scene of varying size and complexity. Speakers delay speech onset if the target object is non-unique and requires disambiguation, and we argue that this reflects the cost of deciding on a high-level strategy for describing it. The eye-tracking data demonstrates that these delays increase when the speaker is able to conduct an extensive early visual search, implying that when a speaker scans too little of the scene early on, they may decide to begin speaking before becoming aware that their description is underspecified. Speakers' content choices reflect the visual makeup of the scene—the number of distractors present and the availability of useful landmarks. Our results highlight the complex role of visual perception in reference production, showing that speakers can make good use of complexity in ways that reflect their visual processing of the scene.

1 Introduction

To describe an object in a visual scene, a speaker must visually analyze the scene, select which of its apparent features and landmarks to mention, and then incorporate the selected content into a linguistic expression. This process requires sensitivity to visual attributes of the target object and of the scene as a whole. Existing evidence from research in psycholinguistics and human vision suggests a tight coupling between speech and vision. Speakers typically look at objects just before mentioning them (Griffin & Bock, 2000) and listeners look quickly at objects they hear (Cooper, 1974; Allopenna, Magnuson, & Tanenhaus, 1998) or expect to hear mentioned (Arnold, Eisenband, Brown-Schmidt, & Trueswell, 2000; Pyykknen & Jrvikivi, 2010). However, prior work has centered on (i) simple scenes, which are convenient for testing targeted manipulations but which underrepresent much of the complexity of real-world scenes (Rubio-Fernández, 2016; Brown-Schmidt & Tanenhaus, 2006; Sedivy, 2003a; Fukumura, van Gompel, Harley, & Pickering, 2011; Pechmann, 1989)), and (ii) analysis of offline production choices, which show speakers' ultimate choices in what to mention but make it difficult to detect moment-by-moment interactions between vision and speech (Sonnenschein, 1985; Koolen, Gatt, Goudbeek, & Krahmer, 2011; Mitchell, van Deemter, & Reiter, 2013a). The classic studies that contributed to the development of Referring Expression Generation (REG) models all used simple visual domains

(maximum 8 objects, typically fewer) and primarily analyzed offline content choices (Pechmann, 1989; Whitehurst, 1976; Ford & Olson, 1975). As such, the focus has not been on how speakers visually understand the scene containing the target they are trying to describe.

Not all studies use offline measures, of course: Analysis of speech onset times can provide a window into the incrementality of speech production (Levelt, 1989; Dell, 1986)). However, the tasks that have measured speech onset rarely probe sufficiently complex visual domains. Likewise, not all studies use simple scenes: Some recent REG work has attempted to move beyond this simplicity by manipulating more fine-grained visual properties like color distinguishability (Viethen, Goudbeek, & Krahmer, 2012) and visual salience (Mitchell et al., 2013a). In our own work (Clarke, Elsner, & Rohde, 2013), REs elicited for targets in complex Where's Wally scenes were found to vary with factors such as scene clutter, the visual salience of a potential landmark, and a landmark's distance to the target, all of which are visual properties posited to play important roles in scene understanding outside of simplified laboratory conditions.

In what ways might visual properties influence production choices and timing? A starting place is to consider factors that are known to confer advantages and disadvantages in visual search and test how such factors influence REG. A new paper in this area (Gatt, Krahmer, Deemter, & van Gompel, 2016) does just this. Search time for a specified target is known to increase as the number of distractor objects increases, but only if the target is distinguished from its distractors by a conjunction of features (e.g., color and size); otherwise the target 'pops out' irrespective of domain size (Wolfe, 2012). (For instance, a lone blue airplane "pops out" of an array of green airplanes; search time is roughly constant regardless of the number of green airplanes.) Gatt et al. report that speech onset times are indeed longer for targets in scenes with more objects just in case the target fails to pop out, presumably due to the time required to check what features should be mentioned to disambiguate the target from the distractors. In contrast, speakers show very short onset times when the target pops out, presumably due to the ease of establishing that color is the only feature needed for disambiguation.

However, Gatt et al.'s claims regarding the role of visual factors in REG are built on experiments using very simple object arrays. Pop-out facilitates REG when all distractors are identical, but that level of simplicity is arguably rare in real-world scenes. Furthermore, a feature that is apparent because it pops out can constitute the entirety of the RE, but only if there is no identical competitor object in the scene. (We distinguish between *distractors*, all non-target objects in the scene, and *competitors*, non-targets which share the basic attributes of the target.) A question therefore is how Gatt et al.'s results scale: How do visual factors influence the incremental production of REs in contexts that increase in complexity and in contexts that do not guarantee the uniqueness of the target object? One hypothesis would be that

whereas pop-out effects emerge for simple scenes and make REG easier, increased visual complexity renders REG harder. This would be due to the time required to find relevant object or scene features that could help disambiguate the target. Alternatively, what may guide processing in complex scenes (as suggested by Clarke et al., see also work on route directions: Klippel & Winter, 2005; Waller & Lippa, 2007; Richter, 2008) is the availability of alternate strategies, such as relative description: speakers could take advantage of easy-to-spot landmarks for inclusion in their REs. A non-unique target, meanwhile, is expected to increase speech planning costs, as posited in REG algorithms that disambiguate a target in a simple domain (see Dale & Reiter, 1995), but what is not known is how production is influenced by speakers' moment-by-moment visual scan to find a competitor in complex scenes.

Thus, REG in complex scenes might represent a compromise between conflicting factors. Visual search of the scene becomes harder as the scene grows more complicated, but formulating an appropriate description might be easier. We argue that speakers navigate the tradeoff flexibly and in a scene-dependent way. If their initial scan of the scene gives them enough information about its large-scale properties, they can often select a descriptive strategy that is likely to be adequate without requiring too much expensive visual search. For instance, rather than searching the whole scene to verify that a possible description uniquely identifies the target, they may constrain the search domain by adding a region phrase like "top left". In a scene with many possible landmarks, they may decide to generate a relative description, while a homogenous scene may favor a strategy using grid coordinates.

Here we report the results of an REG study using scenes that vary in domain size, heterogeneity of the distractors, and the presence of identical competitor objects.[1] We analyze participants' speech onset times and the content of the REs they produce. We also use participants' eye movements to estimate the proportion of the scene participants have viewed in order to further test how speakers' visual understanding of the scene influences REG.

Our results lend support to a model of REG in which vision and speech planning interact, but in ways that reflect the inventiveness of language users, not simply the purported difficulty of visual search. For scene complexity, it is the second alternative proposed above that emerges in the REG results we report here: Alongside a confirmation of the domain size effect (increased onset times for larger scenes) and the pop-out effect (decreased onset times for some scenes with uniform distractors), we find REG facilitation in complex scenes with heterogeneous size/color objects, which likely reflects the availability of salient landmarks which are easy to spot and select during RE planning. Our onset timing findings suggest that in scaling REG models to account for real-world complexity, one need not assume that complexity will always hinder production. Rather speakers can make good use of complexity, in ways that reflect their

[1] Our data and analysis scripts are available at `https://osf.io/9t3x8/?view_only=2ff605ecd63c47b1a7d1e94c1b04124e`.

visual processing of the scene.

Disambiguation, as expected, takes time: We find that the presence of a competitor increases onset time, presumably due to speech planning. However, analysis of onset times alone cannot reveal how speakers' looking behavior influences REG. For a given scene, a speaker may not scan the entire scene before starting to speak. Though the present study uses artificial, controlled stimuli, we believe that these results will continue to apply to real scenes, whose granularity and complexity far exceed the types of scenes typically tested in REG studies. Real-world scenes resist simple feature extraction – one often cannot scan the whole scene to assess what feature dimensions (color, size, shape, material, opacity, orientation, etc.) will be relevant because real-world targets vary in so many, often context-specific, ways.

In our analysis of speakers' eye movements, we find evidence that the proportion of the scene that speakers have viewed influences their sensitivity to the presence of a competitor, a finding that makes sense if one views REG as a tradeoff between the accumulation of (possibly incomplete) information from the visual scene and the utility of saying something rather than nothing. In that light, the simplicity of scenes used in prior work is a special case of the more complicated tasks speakers encounter in the real world.

2 Background and Motivations

Speech planning is known to be incremental (Levelt, 1989): speakers plan utterances online, one phrasal "chunk" at a time (V. S. Ferreira, 1996; F. Ferreira & Swets, 2002), producing filled pauses and revisions that signal delays (Clark & Tree, 2002). In non-visual domains, this planning process is constrained by factors like lexical access and working memory (Levelt, Roelofs, & Meyer, 1999); speakers know the propositions they want to express, but must decide how to realize them in language. In visual domains, however, the semantics of the description reflect facts about the scene itself, including facts of which the speaker might be initially unaware.

To account for this problem of visual access, several studies examine speech production in visual worlds (Tanenhaus, Spivey, Eberhard, & Sedivy, 1995). Such studies provide evidence that production remains incremental in visual domains as well. For instance, speakers look at objects immediately before naming them (Griffin & Bock, 2000) and continue looking at them while they retrieve their names (Meyer, Sleiderink, & Levelt, 1998). The "visual worlds" used in these studies tend to contain small numbers of clearly delineated objects, in part because this allows eyetracked fixations to be reliably resolved to objects in the scene (Fang, Chai, & Ferreira, 2009). In relatively simple visual domains, studies like (Bock, Irwin, Davidson, & Levelt, 2003) conclude that visual processing is fast compared to

linguistic planning, and that it is the latter which accounts for delays in speech onset.

In larger or more complicated scenes, however, visual processing is certainly capable of imposing delays of its own. An extensive literature on visual search (Treisman, 1985; Eckstein, 2011; Wolfe, 2012; Spivey, Tyler, Eberhard, & Tanenhaus, 2001; Reali, Spivey, Tyler, & Terranova, 2006) measures how long it takes listeners to find mentioned objects in an image. Stimuli for visual search experiments can contain large numbers of objects (Spain & Perona, 2010) or use photographic stimuli in which "objects" are not clearly delimited or defined (Wolfe, Võ, Evans, & Greene, 2011; Ehinger, Hidalgo-Sotelo, Torralba, & Oliva, 2009; Clarke, Coco, & Keller, 2013). These studies establish the existence of the "pop-out" effect and explain the conditions under which it occurs. Targets distinguished by the conjunction of two orthogonal features (e.g. color and orientation) do not pop out (Treisman, 1985). In addition, backgrounds which are complex and heterogeneous generally impede visual search (Ariely, 2001; Rosenholtz, Li, & Nakano, 2007; Asher, Tolhurst, Troscianko, & Gilchrist, 2013; Clarke, Green, Chantler, & Emrith, 2008).

In contexts in which REG involves time-consuming visual processing, there are potential effects on both speech onset time and the referring expression itself. Pechmann (1989) is the first to have noticed such effects, finding that some descriptions contain non-canonical adjective orderings (such as "red big"), which he suggests reflect the order in which they were added to the description. In addition, he notes that adjectives which should be prosodically marked as contrastive are sometimes unmarked. Those which contrast with previous descriptions receive stress marking, while those which contrast with descriptors of objects in the current scene do not. He interprets this as another type of disfluency, occurring because the speaker has not yet perceived the entire scene when deciding which elements of the utterance should be stressed.

Brown-Schmidt and Tanenhaus (2006) extend this analysis using eyetracking. They find that, in displays where an adjective (such as size) was contrastive, speakers who saw the contrast object later were more likely to use a postnominal adjective ("the triangle, uh the small one") or a disfluent prenominal adjective ("thee small triangle")[2]. This validates the claim of Pechmann (1989) that these non-canonical orderings reflect the difficulty of visually scanning the scene for a contrasting object and shows more generally that REG depends on the visual properties of the scene and the time it takes for a speaker to scan the scene.

Other visual properties of a scene, especially those known to contribute processing time in visual search, have been shown to influence online production. Above, we reviewed Gatt et al.'s finding that speech onset is slower when the visual system must verify that the target is unique, but fast when the

[2]"thee" represents an elongated production of "the", signaling a disfluency.

target pops out of the scene.[3] Although these results suggest clear consequences when the target pops out of the scene, they have little to say about differences between scenes where the target does not pop out. In scenes like this, the visual component of REG cannot be modeled on analogy with a single visual search. Rather, we must consider visual processing as a set of subtask-specific modules which identify different kinds of visual information which can inform a high-level task (Hayhoe & Ballard, 2014; Tatler, Hayhoe, Land, & Ballard, 2011). A visual search which checks whether a target object is unique may be interleaved with one looking for good landmark objects for use in a relative description, or counting out an object's row-column coordinates on a grid.

Recent work on offline production choices suggests that scene features do indeed have important effects on the content of REs even when scenes are complex enough that targets cannot pop out. For instance, speakers are more likely to use relative descriptions when the scene is more cluttered (in the sense of Rosenholtz et al. (2007)), and are more likely to select an object as a landmark if it is large and visually salient (Clarke, Elsner, & Rohde, 2013). They also use more indefinite determiners in cluttered scenes, acknowledging the potential non-uniqueness of the target in a scene with many objects (Duan, Elsner, & Marneffe, 2013). In addition, when speakers overspecify the description of a landmark object, they are likely to add a visually distinctive property (Paraboni, Galindo, & Iacovelli, 2015). In monochrome displays in which the target is the same color as the distractors, speakers produce fewer redundant adjectives than in chromatic displays in which color provides a useful cue for visual search (Rubio-Fernández, 2016). Results of scene properties like feature redundancy show that RE content is affected by properties of the entire stimulus which go beyond target pop-out.

While Rubio-Fernández does not directly address the time course of RE planning, her comparison between English speakers and Spanish speakers also provides some information about the relative timing in production of color adjectives. Her results show that if color information can be provided early in the utterance, presumably to guide visual search, speakers are more likely to opt to include a color adjective. Specifically, English speakers were found to produce more redundant color adjectives than Spanish speakers, which she attributes to the fact that adjectives in English appear pre-nominally ('*blue* cup') and hence offer listeners early information about a visually salient property of the target, whereas adjectives in Spanish appear post-nominally ('taza *azul*') and hence provide a less efficient contribution to the listener's visual search. The hypothesis explored here is that visual effects are directly relevant to the online time course of RE planning.

[3] In accord with their hypothesis, uniquely colored targets have small, constant speech onset times. For targets identifiable only by a combination of size and color, speech onset times increase with the number of distractor objects. This is expected, since this is a visually difficult conjunction search which requires a sequential scan of the entire scene to find a possible competitor if one exists. Gatt et al. predicted that uniquely-sized targets would also pop out, leading to constant speech onset times. This was not the case, suggesting that uniquely-sized objects may not benefit from the pop-out effect.

Examining REG in visually complex scenes also has consequences for the design and interpretation of computational REG models (Krahmer & van Deemter, 2012). Many such models are variants of the Incremental Algorithm (IA) (Dale & Reiter, 1995), in which a description is built up by iterating over possible distractor objects in the scene and, if the current description fails to distinguish the target from a distractor, adding new content until the target is fully disambiguated. While the IA and its descendants are all incremental in the sense that content is added piece-by-piece, they rarely specify when content planning is taken to occur relative to speech itself.

For instance, the Visible Objects Algorithm (Mitchell, van Deemter, & Reiter, 2013b) makes immediate selections for color and other perceptible non-relative properties of the target before making an IA-like scan of the distractors. While these models acknowledge the importance of visual processing to REG, they leave unspecified when this visual processing takes place relative to the actual speech. One can interpret modeling decisions like these as a reflection of early planning (color is added to the selected content first, but speech begins only after the scan phase) or a consequence of planning after speech onset (the model "starts speaking" immediately and utters the color term immediately after deciding to include it).

Since these models do not specify the timing of REG, they cannot be taken as fully-specified psycholinguistic models, a point also made in Gatt et al. (2016). More importantly for designers of computational REG algorithms, the timing of visual search can affect the *form* of the description. As already described, timing is important in predicting the presence of speech disfluencies (Brown-Schmidt & Tanenhaus, 2006) and focus intonation (Pechmann, 1989). It might also influence the order in which landmarks and targets are mentioned in relative descriptions (Clarke, Elsner, & Rohde, 2015). Thus, computational models which capture the role of real-time planning on REG may be able to produce more human-like prosody and syntax.

3 Experimental design and predictions

To test our hypothesis that different kinds of visual complexity can alternatively facilitate or hinder REG, we ask speakers to describe target objects in a variety of visual scenes. Figure 1 shows two sample scenes. The experiment uses a $4 \times 3 \times 2$ design that manipulates the number of objects in the scene (approximately 25, 49, 81, 121), the visual heterogeneity of those objects (uniform, multicolor, skewed), and the status of the target as unique or non-unique. Of interest is evidence for how these scene properties influence speakers' efforts to visually understand the scene and to plan their speech. We analyze their eye movements, speech onset times, and their ultimate RE productions.

Our manipulation of grid size substantially increases the number of objects in the scene compared

with most previous visual world studies and controlled REG studies, which typically contain fewer than a dozen objects, often no more than 4. With larger numbers of objects, the scene is by definition more complex though pop-out effects may eliminate REG difficulties (as demonstrated by (Gatt et al., 2016) on stimuli containing up to to 16 objects). At least two different visual strategies are possible in response to this complexity. On one hand, speakers may sequentially search through the objects in the scene, likely using a mix of left-to-right (Gilchrist & Harvey, 2006) and stochastic (Clarke, Green, Chantler, & Hunt, 2016) scanning strategies. In this case, if the impact of visual factors on REG is analogous to the impact of such factors on visual search, more objects could yield longer scan times and hence delayed onset times. Alternatively, speakers may avoid a whole-scene scan by restricting the relevant domain with salient landmarks or terms like "near the top right". If that is the case, speakers' visual processing of the scene could treat many of the objects as undifferentiated "background" or "texture" (Landy & Bergen, 1991), rather than requiring an explicit inventory of all objects' attributes.

Given the possibility of interpreting distractor objects as "background", we explicitly manipulate the heterogeneity of those distractors. If speakers process the growing number of objects in the scene as background, we expect them to make use of visually salient non-target "foreground" objects as potential landmarks. Scenes with such objects will necessarily be more complex and visually cluttered, but also offer more opportunity to use relative descriptions (Viethen & Dale, 2011; Clarke, Elsner, & Rohde, 2013). We use three levels of scene hetereogeneity: *uniform* scenes in which all objects are identical in size and color, varying only in shape, *multicolor* scenes in which objects vary in shape and color, and *skewed* scenes, in which a small number of objects are larger, brighter, or uncommonly colored. Although the visual search literature suggests that skewed and multicolor scenes impose a larger burden on visual processing (i.e., finding "the green circle" may be challenging) REG may be easier in such scenes due to the availability of landmarks for relative descriptions.

Lastly, the manipulation of target uniqueness is intended to create a task that approximates another aspect of the complexity of real-world tasks: In describing a real-world target to an interlocutor, a speaker must determine whether their description will uniquely pick out a single object or whether it will be ambiguous. When a target in our experiment is non-unique, the scene contains one or more competitors which are identical to the target in all attributes (color, shape and size). In that case, a simple description like "the green circle" is ambiguous. If speakers are sensitive to the presence of a competitor, they will need to use disambiguation strategies which go beyond naming the attributes of the target object itself. They may use region terms ("top left"), landmarks ("next to the") or grid coordinates ("two down"). This range of possibilities reflects the diversity of REs elicited in more realistic tasks such as navigation (Curry, Gkatzia, & Rieser, 2015).

Our manipulation of target uniqueness contrasts with Gatt et al.'s (2016) study, in which all trials contained unique targets (including filler trials). This may have influenced participants' strategies if over the course of the experiment they came to realize that no identical competitor would ever be present and hence no scene-relative descriptor would ever be needed. We expect that our participants will produce REs that show a wider variety of linguistic strategies. These strategies are particularly important because speakers might use them in preference to scanning the entire scene— by including a region term like "top left", speakers can save themselves the visual effort of checking for a competitor in the bottom right. Thus, we expect differences between REs and onset times when the speaker has quickly verified the presence of a competitor and when they have not (Brown-Schmidt & Tanenhaus, 2006).

Considering our three dependent variables, the content of REs that speakers produce are predicted to vary with the presence/absence of a competitor: Non-unique targets should yield more complex REs than unique targets because disambiguation is non-trivial. In addition, heterogeneity is predicted to guide RE strategies, particularly when the target requires disambiguation from a competitor: Skewed scenes provide more salient landmarks for mention in an RE, while uniform scenes permit REs based on visual pop-out ("the only green circle") if there is no competitor. As our results show, these predictions are upheld, and speakers' productions point to the variety and creativity of possible RE strategies.

For speech onset time, the increased planning for descriptions of non-unique targets is predicted to cause delays compared to those for unique targets. Larger scenes are likely to delay speech onset time, as found in previous work (Gatt et al., 2016), because the visual effort of scanning for competitors increases with the number of objects. In more heterogeneous scenes, however, onset times are subject to two conflicting pressures. The increased effort of visually searching for a competitor will tend to increase onset times, but the increased availability of landmarks for relative descriptions may make planning easier. Our results suggest that both pressures are active. Visual search cost dominates in scenes with unique targets, where the resulting RE itself is simple but the visual search to confirm the lack of competitor becomes more difficult with increasing complexity of the grid. Wwhen the target is non-unique, more effort is needed to construct a suitable disambiguating RE, and therefore scene complexity yields a facilitatory effect from the availability of landmarks in more heterogeneous conditions.

Lastly, speakers' eye movements provide information about how much of the scene they have viewed and how this impacts speech onset: Do onset times reflect a distinction between cases in which a speaker's early scan covers more of the scene, likely confirming target uniqueness early, versus cases in which the uniqueness status of the target is unknown due to more limited viewing in the early scan? Brown-Schmidt and Tanenhaus (2006) found that these such cases differ in fluency and order of mention. We expect also to find effects on speech onset time: Scenes with non-unique targets have delayed onset

times relative to unique targets—but this gap is expected to narrow when the early visual scan does not cover much of the scene. As expected, our results show that if the speaker cannot quickly verify which condition they are in, they are unable to modify their onset time in a condition-appropriate way.

To reiterate the predictions:

- RE content: The presence of competitors should lead to descriptions with more diverse referential strategies (landmarks, relative phrases, grid coordinates); the selection of particular strategies should vary with scene size and complexity.
- Onset time: Competitor objects should slow speech onset time; larger scenes should slow speech onset time; onset times should vary by scene type (the expected effects are unclear a priori).
- Eye movements: More extensive visual scanning early in a trial should allow participants to establish competitor presence, which is assumed to influence RE planning time and in turn onset time; less visual scanning early in the trial should reduce the slowdown caused by competitor presence, making onset times for trials with competitors more similar to those without.

4 Methods

Participants

Eighteen native English speakers (11 female, ages 18-38, mean age 23.8) from the University of Edinburgh student population were recruited and received £10 for their participation in the study. All subjects self-identified as having normal or corrected-to-normal vision. Data from an additional two participants was not analyzed due to a recording error.

Materials

Visual scenes were pseudo-randomly generated, each consisting of a square grid of colored squares and circles (see Figure 1). As described in the previous section, stimuli varied along three visual dimensions: *heterogeneity*, *size* and *presence of competitors*. Heterogeneity affected the color and size distribution of the shapes in the grid. In *uniform* stimuli, the shapes were all small and all the same color. In *multicolor* stimuli, the shapes were randomly colored red, blue or green in equal proportion. In *skewed* stimuli, shapes were randomly colored red, blue, green, or bright red, bright blue or bright green, but in unequal proportions: approximately 76% of the shapes were assigned to the most common color, with other colors being represented 15%, 9% respectively. 10% of the shapes were set to a brighter version of their color. 10% of shapes were also made twice as large as standard for that grid size.[4] The

[4]In radius; one large shape occupies roughly the area of four small ones.

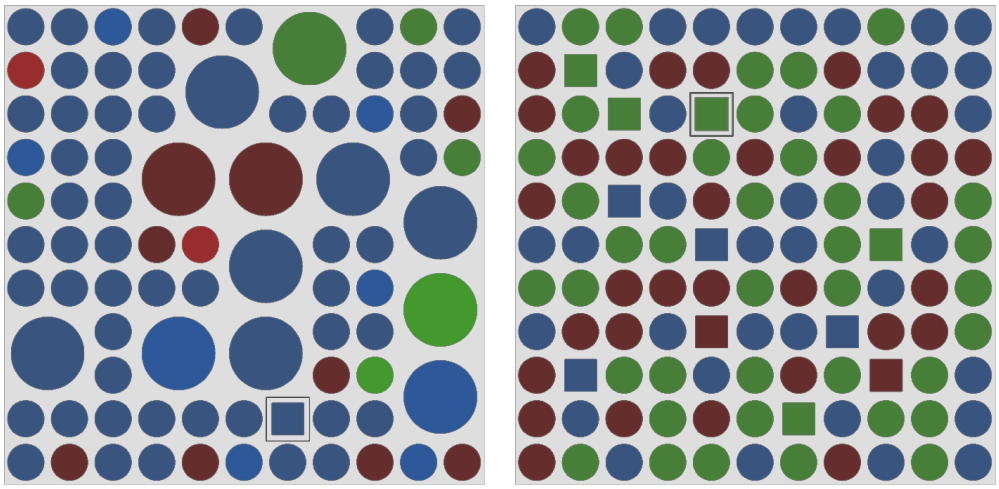

Figure 1: Two sample experiment materials with target indicated via a box. Both grids show the largest grid size (11×11). The grid on the left shows a target that is unique, in a *skewed* distribution of size and color. The grid on the right shows a target with same-shape competitors in the *multicolor* condition.

grid size of a stimulus is measured as the maximum number of small objects in a row or column of the grid ($N = 5, 7, 9, 11$) or the number of cells in the grid, N^2. In the skewed condition, the presence of large objects means that there are fewer than N objects in some rows and fewer than N^2 objects in the grid.

In each stimulus, a particular object was designated as the *target* of the description. In the *no competitor* condition, the target was the only object of its shape—either the only square among circles or the only circle among squares. In the *competitor* condition, other objects had the same shape as the target, and in particular, there was always at least one object the same shape and color as the target. Competitor trials were created by sampling 10% of the initially generated objects and altering their shape and color to match the target, so long as they were viable competitor candidates (not already the target, or large/bright/edge/center objects). The sampling was with replacement[5]; thus, stimuli of a given size had a variable number of competitors. Targets and competitors were never large or bright objects, never located along the edges of the grid, and never the center object.

Procedure

Participants were recorded in a booth, with the experimenter present but behind a window; participants could both hear and speak to the experimenter through the glass. There was no interlocutor, although they were instructed to imagine a scenario in which they were speaking to another person, and might have considered the experimenter in that light. Stimuli were displayed on a 24" BENQ monitor,

[5]Due to a software bug. This meant that some trials intended to have competitors did not; see below.

with participants viewing the scenes from roughly 22" away.

Each participant viewed 120 scenes, with no filler or practice trials. Participants were instructed to describe the target object so that another person viewing the scene would be able to pick out that object. They were encouraged to produce their description quickly and accurately. These instructions were intended to be comparable to those from (Gatt, van Gompel, Krahmer, & van Deemter, 2012), who instructed participants to "speak naturally and clearly, but respond as quickly as possible given the conditions", and to encourage speakers to focus on the descriptive task. At the start of each trial, the participant was presented with a blank screen with a fixation cross at the location where the target would appear. Thus, participants were not required to visually search for the target before beginning to plan their utterance (as was the case in Gatt et al. (2016)'s study). The appearance of the grid was dependent on the participant looking directly at the fixation cross, at which point the experimenter revealed the full grid. If the participants' fixation was sufficiently far from the fixation cross, the experimenter initiated a re-calibration process.

The full scene appeared, along with a "beep" sound, and the participant then gave their description, treating as their target whatever object appeared at the location where the fixation cross had been. Participants' descriptions were recorded using an Audio-Technica AT8531 microphone worn around the neck, and their eye movements tracked with a head-mounted eye-tracker (SR Research Eye-link 2). No data was excluded on the basis of tracker errors.

Each participant saw a different unique set of stimuli in random order. All stimuli were displayed at the same actual size and orientation, regardless of the number of objects in the grid. (Thus, the radius in pixels of an object is smaller for grids with more objects.) The complete set of stimulus images, the generator script, and the transcribed recordings are available at `https://osf.io/9t3x8/?view_only=2ff605ecd63c47b1a7d1e94c1b04124e`.

Extraction of RE content and timing

We used a combination of manual and automated techniques to prepare the raw data for analysis. The participants' recorded utterances were manually transcribed and then aligned with the recordings using the Penn Forced Aligner (Yuan & Liberman, 2008). Speech onset times and word durations were extracted directly from the aligner output. Onset is determined by the time from presentation of stimulus to the beginning of the first word (disfluencies like *uh* and *um* are not counted as speech onset).

For RE content, we automated the identification of participants' descriptive strategies. We assume that a description is made up of *descriptive elements*, each carrying information about some particular aspect of the scene. In this, we followed roughly the taxonomy used in Clarke, Elsner, and Rohde (2013), which distinguishes between landmark, target and region phrases. We added categories for coordinate

and scene-relative descriptions, since our participants often used these. We distinguish the following types of descriptive elements (examples in parentheses):

Target Non-relative reference to the target object, possibly with some of its attributes (shape, color, size and brightness[6]) (*a green block*)

Scene-relative A description relative to the overall distribution of shapes in the scene (*the only circle; the lowest blue circle*)

Region Reference to a part of the scene (*at the upper left*)

Coordinate A grid reference by row or column (*the fourth row from the bottom; the second shape across*)

Landmark Reference to some other object near the target (possibly with attributes) (*above the red circle; surrounded by blue squares*)

Other Anything else (*I'm looking at; we're looking for a; it's the*)

We heuristically labeled these elements using a series of hand-written deterministic rules that rely on key lexical items; for instance, a preposition followed by an object description was tagged as a landmark, whereas numbers and the words "row" and "column" signal coordinates. See Appendix 9 for the complete list, and its accompanying Figure 7 for examples of tagged utterances.

For the analysis of RE content, we test whether participants' use of these different descriptive elements varies with properties of the scene (section 5.1). For the analysis of speech onset times, we test whether participants' time to start speaking varies by condition (section 5.2).

Eye data

Gaze data were processed to fixations using the default SR research Fixation detection algorithm. In a post-hoc analysis, we use these fixations to test how speech onset times vary with how much of the scene participants have scanned (section 5.3).

5 Results

There are 2046 complete recordings, 853 with competitors and 1193 without. The imbalance between conditions was due to a bug in the stimulus generation script[7]. The distribution of different hetero-

[6]It is possible that an attribute like size or brightness could be a relative descriptor because its computation requires inspection of other objects, but in this experiment, objects appeared only in two possible sizes and two possible brightness levels. Given this predictability across scenes, we treat these two attributes as *Target* descriptors which can be assessed via inspection of the target only, akin to shape and color; that is, size is an intrinsic property of the object (Sedivy, 2003b).

[7]In some cases, the target and competitor were selected to be the same item in the search array. This effectively increased the number of trials in the no-competitor condition.

- It's the circle
- It's a blue circle in the top right hand side the third row from the top and the third column from the left
- A green square
- Um I'm looking at um a screen completely filled with circular shapes bar one um I'm looking at the shape that is is not a circle
- Okay we're looking for a small blue square it's the only square present on the screen it's around the middle of the screen
- The red circle
- It's a bright green square next to a red circle in the top left hand sides of the screen

Figure 2: Sample participant referring expressions from the corpus.

geneities and grid sizes was balanced. 114 trials were discarded because, due to recording problems, they did not produce a valid recording (59), the recording could not be completely transcribed (32), or it could not be aligned (23). R scripts and dataframes for the analyses described below are available at `https://osf.io/9t3x8/?view_only=2ff605ecd63c47b1a7d1e94c1b04124e`. A few selected referring expressions are shown as Figure 2.

5.1 RE content

Our first planned analysis tests the impact of our experimental manipulations on referring expression *content*. This is intended to replicate previous analyses in showing that, in our stimuli, adding a competitor object does in fact lead speakers to construct referring expressions that contain additional disambiguation information (Engelhardt, Bailey, & Ferreira, 2006). We also check that the use of *landmark* phrases (relative descriptions) is linked to scene heterogeneity— speakers are predicted to use more relative descriptions when the scene contains highly salient distractor objects (as was the case in the complex scenes in (Clarke, Elsner, & Rohde, 2013)). In the remainder of the paper, we italicize references to descriptive elements in order to distinguish them from normal usage of the words: *landmark* phrase versus visual landmark.

Table 1 shows how often speakers' REs include include an element of each type across the competitor-present and competitor-absent conditions. As expected, participants use more *region*, *coordinate* and *landmark* when a competitor is present, whereas *scene-relative* elements such as "only" are more common in scenes without a competitor. The descriptive elements *target* and *other* do not vary strongly with the presence/absence of a competitor and are both relatively frequent, showing that most participants

Table 1: RE content: Percentage of descriptions including an element of each type, by competitor presence

	Target	Scene-Rel.	Region	Coord	Lmark	Other	Total Descs.
Competitor	90%	25%	60%	39%	40%	83%	853
No comp.	94%	35%	39%	15%	19%	78%	1193

include elements of these types regardless of the presence of a competitor. The results indicate that our competitor manipulation is successful; participants use *region*, *coordinate* and *landmark* elements to disambiguate targets from competitors. This implies that scenes with competitors require more content planning effort than those without. (The proportions of color, size and shape terms in descriptions are attached as Appendix 10.)

We construct logistic mixed-effects models to predict the presence or absence of each descriptive element in an RE, given the scene features. We use fixed effects for grid size, competitor presence and scene heterogeneity along with all interaction terms. Hetereogeneity is difference coded using two real-valued variables, indicating the differences as the condition changes from skew to multicolor and from multicolor to uniform. We use an uncorrelated random slope and intercept for each fixed effect by speaker, as recommended in (Barr, Levy, Scheepers, & Tily, 2013).[8] Grid size is z-transformed; competitor is difference coded as -1 (absent) or 1 (present). Models are fit using LME4 1.17 (Bates, Maechler, Bolker, & Walker, 2014) and OPTIMX/LBFGS or BOBYQA. Convergence is assessed by checking that $\max|grad| < 1e-4$. Each fixed effect is tested for significance by using ANOVA to compare a model which lacks that fixed effect to the full model, with threshold $p < .05$; these individual tests indicate the conclusion we would reach if we had selected only a single hypothesis for testing *a priori* (note that the maximal model structure contains many interactions which we do not consider interesting *a priori*, but we add and test these for methodological reasons). To account for multiple comparisons, we also report which factors are significant when controlling the false discovery rate (Benjamini & Hochberg, 1995) (the ratio of falsely rejected null hypotheses to total rejections) so that $FDR < .05$, using a procedure designed for correlated statistics (Benjamini & Yekutieli, 2001).

Coefficients are shown in Table 2. For the 3-way heterogeneity factor, Table 2 uses the $\Rightarrow$ notation to indicate difference coding. For example, the coefficient for *Multi⇒uniform* shows the difference in probability of a given phrase type between the multicolor condition and the uniform condition.

The mixed-effects model supports the basic patterns observed in Table 1. Competitor presence significantly encourages the use of *region*, *coordinate* and *landmark* phrases ($\beta = 1.26, 1.38, 1.30$).

In addition, the results confirm that RE content varies with other scene attributes in ways that reflect the utility of particular descriptive strategies for certain scenes. Speakers' use of *regions* varies not

[8] See Table 1, model 6 (Barr et al., 2013).

Table 2: RE content analysis: Effects from mixed-effect binomial models predicting phrase type inclusion[a]

Effect ***(std. err.)***	**Target**		**Scene-rel**		**Region**		**Coord**		**Lmark**	
Intercept	4.88	⁎⁎⁎	-1.98	⁎⁎	-0.46		-2.42	⁎⁎	-2.19	⁎⁎
	(0.72)	•	*(0.57)*		*(0.93)*		*(0.74)*		*(0.65)*	
Main effects										
Competitor	0.18		-0.27		1.26	⁎⁎⁎	1.38	⁎⁎⁎	1.30	⁎⁎⁎
	(0.54)		*(0.41)*		*(0.30)*	•	*(0.29)*	•	*(0.22)*	•
Grid size (z-trans)	0.28	†	0.15	†	0.39	⁎⁎⁎	0.28	⁎	-0.01	
	(0.16)		*(0.08)*		*(0.08)*	•	*(0.11)*		*(0.12)*	
Multi⇒uniform	-0.39		0.36	⁎	-0.26		-0.02		-0.52	†
	(0.32)		*(0.18)*		*(0.24)*		*(0.22)*		*(0.29)*	
Skew⇒multi	-0.23		-0.07		0.01		1.00	⁎⁎	-0.79	⁎⁎
	(0.41)		*(0.20)*		*(0.18)*		*(0.32)*		*(0.20)*	
2-way interactions										
Competitor × grid size	0.11		-0.03		-0.01		0.22	⁎	0.17	
	(0.14)		*(0.07)*		*(0.08)*		*(0.09)*		*(0.11)*	
Competitor × multi⇒uniform	0.42		-0.06		0.27		0.87	⁎⁎⁎	0.20	
	(0.30)		*(0.18)*		*(0.26)*		*(0.20)*	•	*(0.28)*	
Competitor × skew⇒multi	0.14		0.12		0.03		-0.17		0.44	⁎
	(0.39)		*(0.19)*		*(0.18)*		*(0.22)*		*(0.19)*	
Grid size × multi⇒uniform	-0.62	⁎	0.16		0.23		-0.05		0.42	†
	(0.31)		*(0.18)*		*(0.20)*		*(0.21)*		*(0.21)*	
Grid size × skew⇒multi	0.40		-0.05		-0.16		-0.06		-0.40	†
	(0.30)		*(0.18)*		*(0.19)*		*(0.21)*		*(0.21)*	
3-way interactions										
Competitor × grid size × multi⇒uniform	-0.15		-0.12		0.13		0.22		0.02	
	(0.35)		*(0.21)*		*(0.22)*		*(0.20)*		*(0.23)*	
Competitor × grid size × skew⇒multi	0.18		0.10		0.10		-0.37		0.24	
	(0.30)		*(0.18)*		*(0.19)*		*(0.25)*		*(0.19)*	

[a]†: $p < .1$, ⁎: $p < .05$, ⁎⁎: $p < .01$, ⁎⁎⁎: $p < .001$; •: part of set R with expected $FDR(R) < .05$

only with competitor presence, but also with grid size ($\beta = 0.39$): Larger grids encourage more *region* phrases, as one would expect since mentioning an area like "top left" has the power to eliminate more distractors in a larger array.

Speakers' use of *coordinates* varies with scene heterogeneity: *coordinates* are used more often in multicolor and uniform scenes than in skewed ones ($\beta = 1.00$ for skew to multi, non-significant between multi and uniform). This may reflect the fact that the objects in the non-skewed scenes more obviously align to the grid, making counting ("third column, fourth one down") more attractive.

The remaining trends are non-significant when controlling FDR at the 0.05 level, and should be interpreted with caution; on average, we should expect one or two false positives in this set. However, they are generally interpretable and in accord with our predictions.

Coordinate usage increases with grid size ($\beta = 0.28$). Multicolor and uniform scenes yield different rates of *coordinates* when a competitor is present: *coordinate* use increases in uniform scenes when a competitor is present (competitor $\times$ multi to uniform interaction, $\beta = 0.87$). This presumably reflects the fact that identifying the target object cannot be accomplished by merely listing the target attributes and hence a favored strategy for this scene type is recruited. The effects of grid size and competitor are likewise not simply additive, with large grids containing a competitor yielding increased use of *coordinates*. This may be due, as (Gatt et al., 2016) suggests, to the increasing difficulty of REG in larger arrays causing speakers to prefer to disambiguate more.

For *landmarks*, heterogeneity has the predicted effect on the use of relative descriptions: REs in skewed scenes use more *landmarks* ($\beta = -0.79$) than multicolor. These effects are opposed to those for *coordinates*, which suggests that the two strategies are used in complementary distribution; skew$\Rightarrow$multi yields more *coordinates* ($\beta = 1.00$) and fewer *landmarks* ($\beta = -0.79$). We proposed above that skewed scenes would encourage speakers to mention landmarks in their descriptions, and that the availability of this strategy might speed up their REG planning. The result here confirms the first (linking) part of that hypothesis.

The effects of competitor and scene type on *landmark* inclusion are not quite additive, as shown by the significant 2-way interaction: The strong dispreference for using the *landmark* strategy in uniform scenes is partially reversed if a competitor is present (competitor:multi to uniform, $\beta = 0.44$), even though few landmarks are available. In general, *landmark* descriptions used for uniform scenes are complex and refer to the locations of multiple shapes: "It's a blue circle in the left— the right hand side, second column from the right on the centre row **surrounded by blue squares**"; "the blue square in the middle row, three rows from the top; on that row it is also the most right hand square **of the three squares**". This interaction is unexpected— we failed to anticipate the range of landmark types

	Unique targets	Non-unique targets
Total	1193	853
Contain color	1011 (85%)	766 (90%)
Contain size	138 (12%)	112 (13%)
No color or size (not overspecified)	181 (15%)	-
No relative info (underspecified)	-	177 (21%)
Filled pause or partial word	286 (24%)	303 (36%)

Table 3: Descriptive statistics on description over- and underspecification.

participants would consider— but is easy to interpret given the data. It is relatively weak in comparison to the main effects.

Target and *scene-relative* descriptions vary little between conditions. Relatives are more likely in uniform grids ($\beta = 0.36$), perhaps because unique shapes ("the **uppermost** top left blue square") are more likely to pop out.

Target descriptions are slightly less likely in large, uniform grids ($\beta = -0.62$), an unexpected effect which might occur because the similarity of all the objects reduces the number of salient attributes a participant might consider mentioning (shape, color, size, brightness). *Scene relatives* are slightly more likely in uniform grids ($\beta = 0.36$), perhaps because unique shapes ("the only square") are more likely to pop out.

The lack of a difference in the use of *scene relatives* between competitor present and competitor absent conditions is unexpected. The most common *relative* in our corpus is "only", which is harder to use if the target has a competitor. But speakers do find ways to employ it for non-unique targets, for example by qualifying the domain ("there is only one pair of vertically adjacent squares and it is the lower of the two").

Although over vs. underspecification is not the focus of this work, we conduct a brief post-hoc analysis of description content using the automatically assigned lexical categories in order to assess the adequacy of the descriptions provided (Table 3). For unique targets, we check for overspecification (any attributes beyond shape). Nearly all descriptions of unique items contain a color term, indicating widespread overspecification of color (Koolen et al., 2011; Pechmann, 1989). Relatively few descriptions contain size terms, and many of these are potentially part of landmark descriptions, since the target objects never vary in size.

For targets with competitors, we cannot automatically check for overspecification, but can detect some cases of underspecification by looking for descriptions without relative information (coordinates, prepositions or region information). 21% of the descriptions of targets without competitors were underspecified (that is, the speaker produced a description like *it's the blue circle* for a scene with multiple

blue circles). Visual search for a competitor object can be difficult; while some of the underspecified cases might reflect task fatigue, many of them are presumably cases where the speaker did not detect the competitor before beginning to speak. While more than half the participants (10 of the 18) produced at least one underspecified description (according to these criteria), some participants were more prone to this form of carelessness than others. 66% of these descriptions were produced by only 3 speakers.

In addition to examining the content of the final description, several previous studies have used hesitations and non-canonical ordering of descriptive elements as evidence about the production process. While a full analysis of these phenomena is beyond the scope of this study, it is possible to automatically detect a few specific cases studied in this previous work. Only 28 utterances have color adjectives immediately followed by size adjectives, the non-canonical ordering studied by Pechmann (1989). These are utterances like: "next to the bright big bright red square", which also seems to incorporate a revision. More complicated cases of post-nominal adjectives ("the square— the big one") studied by Brown-Schmidt and Tanenhaus (2006), are more common in the corpus, but difficult to automatically detect without parsing the utterances syntactically, since the added information could be the beginning of a new clause referring to a different object.

Disfluencies and pauses, also studied by Brown-Schmidt and Tanenhaus (2006), are also complicated to detect. It is possible to gain some idea of the relative proportions of revised descriptions by counting filled pauses and partial words in the transcripts (such as "okay, blue square, uh, s[quare], bo[ttom], bottom row"). We find that descriptions in both conditions frequently contain these (Table 3), but that descriptions of non-unique targets are more likely to contain them (36%) than descriptions of non-unique targets (24%). These measurements are sufficient to support the finding of Brown-Schmidt and Tanenhaus (2006) that the presence of a competitor reduces fluency. However, we do not attempt to use these indicators of disfluency in our analysis of speech onset times because they miss a large number of revisions and disfluencies which contain unfilled pauses or phrasal repairs without partial words (Shriberg, 1994). For instance, this utterance from the corpus is disfluent (reparandum in italics, repair in bold), but would not be detected by looking for partial words: "on the left of the screen is *a line* **a vertical line** of three circles. . . ." We leave detailed annotation and investigation of utterances like this for future work.

5.2 Speech onset times

Here, we examine speech onset time as an indicator of the difficulty of planning an RE. In the previous section, we found that the presence of a competitor object leads speakers to adopt more complex descriptive strategies (*regions*, *coordinates* and *landmarks*), and that the skewed visual condition lead to an increased reliance on landmarks. Here, we show that these tendencies are reflected in the online timing

of REG.

We measure the time to speech onset as the interval between the first presentation of the scene to the participant and the beginning of the first word. On average, onset to speech is 2.1 seconds long, with a standard deviation of 1.4 seconds. This is enough time for participants to make several visual fixations (mean 6.0, median 5),[9] although not enough to permit an exhaustive examination of every object in the scene.

Following Gatt et al. (2016), we predict grid size to increase onset times since planning becomes more difficult with more distractor objects to scan. The presence of a competitor should also increase onset times, since in this case, the speaker typically plans a more complex RE with extra disambiguating information. Increased scene heterogeneity might have effects in both directions. A skewed scene is harder to visually search, but encourages the use of landmarks, which may be easier to plan than alternate strategies like counting out row/column coordinates.

We test these predictions by constructing a linear mixed-effects model. We model log onset time[10] as a function of grid size, competitor presence and visual condition along with all two-way interactions. As before, we use uncorrelated random slopes and intercepts per participant.[11] Fixed effects are tested for significance using leave-one-out ANOVAs with threshold $p < .05$; effects are shown in Table 4. Figure 3 shows the onset times as a function of grid size, heterogeneity, and presence of a competitor.

The results confirm the predicted effects of grid size ($\beta = 0.04$) and competitor ($\beta = 0.12$), with larger grids and grids with a competitor yielding longer onset times. The effects of scene heterogeneity are not significant on their own, but heterogeneity interacts with competitor: uniform scenes with a competitor yield longer onset times ($\beta = 0.07$; this trend is non-significant when controlling FDR). This latter interaction accords with our prediction that some types of scene complexity can facilitate REG planning. We attribute this pattern of results to the differential use of landmarks: Speakers in uniform scenes use fewer landmarks, while in skewed scenes they use more; if noticing an easy-to-spot landmark can speed RE planning, the increased complexity of skewed scenes need not hinder production and in fact, as we show, can reduce speech onset times.

These effects are apparent in Figure 3. The largest effect, the delay caused by competitor presence, is represented by the gap between the dashed and solid lines. The effect of grid size is apparent as a general upward trend in most conditions. The ordering of the three solid lines shows a (non-significant) trend whereby increased visual complexity yields speech onset delays. In the dashed (competitor) group, the upward slope of the purple (uniform) line corresponds to a significant interaction: Uniform scenes might

[9] Somewhat more than Pechmann (1989), who finds a mean of 3.29 fixations before onset, perhaps due to the increased visual complexity of the scenes.

[10] Onsets are right-skewed (3rd quartile 2.4, max 18.6), so we log-transform them to reduce the influence of outliers.

[11] Models are fit using LME4 1.17 and the default BOBYQA optimizer.

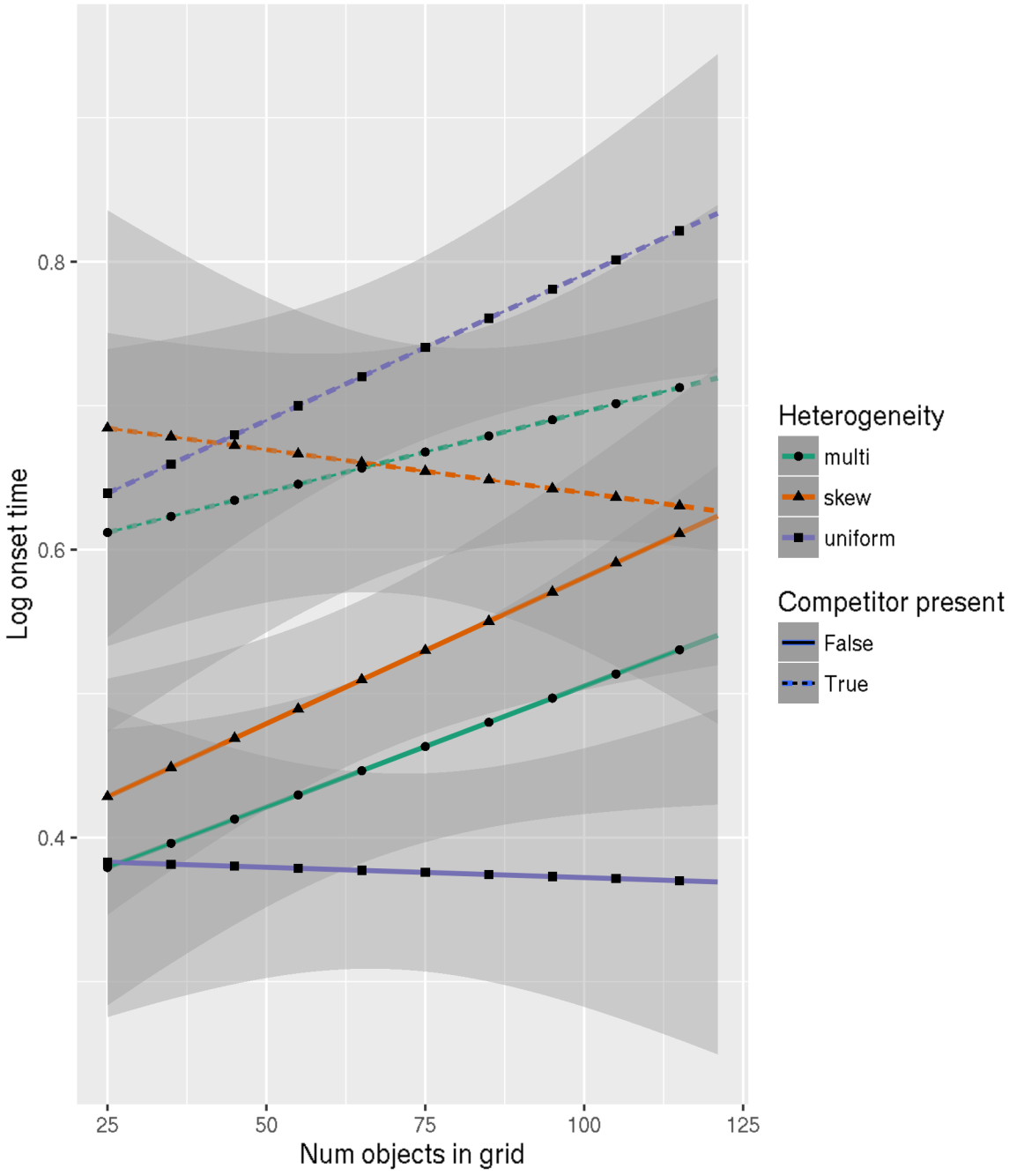

Figure 3: Log onset time as a function of grid size, grouped by heterogeneity and presence of a competitor

Table 4: Predictors in mixed-effects linear model of log onset time[a]

Effect	β		std. error	t value
Intercept	0.56	$***$ •	0.09	6.62
Competitor present	0.12	$***$ •	0.03	4.21
Grid size (z-trans.)	0.04	$**$ •	0.01	3.12
Multi⇒uniform	-0.01		0.03	-0.51
Skew⇒multi	-0.01		0.03	-0.32
Competitor × grid size	-0.01		0.01	-0.67
Competitor × multi⇒uniform	0.07	$*$	0.03	2.43
Competitor × skew⇒multi	0.04		0.05	0.73
Grid size × multi⇒uniform	-0.01		0.03	-0.32
Grid size × skew⇒multi	0.01		0.03	0.47
Competitor × grid size × multi⇒uniform	0.05	†	0.03	1.73
Competitor × grid size × skew⇒multi	0.04		0.05	-.83

[a]†: $p < .1$, $*$: $p < .05$, $**$: $p < .01$, $***$: $p < .001$; •: part of set R with expected $FDR(R) < .05$

be expected to correspond to the easiest visual searches, but when more disambiguation is required, they appear to demand additional effort. The downward slope of the blue (skewed) line is not significantly different than the red multicolor line, but again trends in the direction of complexity facilitating REG.

We conduct a brief exploratory analysis of the effect of underspecification (as defined in subsection 5.1) on onset times. Since most of the detectably underspecified descriptions were produced by only a few speakers, we cannot precisely quantify the correlation between them, but our preliminary analysis does suggest a connection. The onset times of the 177 underspecified descriptions (descriptions of non-unique targets which lacked any relative information) averaged 2.03 seconds (dev 1.67) while the onset times for descriptions of non-unique targets with relative information averaged 2.46 seconds (dev 1.50). The implication is again that failure to find the competitor object early in visual search results in a simple description and a quick onset time. Our overall analysis applies to all cases where the competitor object is not found before speech onset. Most underspecified descriptions are presumably produced in the subset of these cases where the speaker never finds the competitor at all.

5.3 Scan distance

In the section above, we show that speech onset slows down when the REG planning problem is more difficult, either visually or linguistically. Here, we use participants' eye movements to analyze how their speech onset times vary with their trial-by-trial looking behavior. Speakers do not scan an entire scene before starting to speak, and the REs they eventually produce depend on what scene features they can discover early on (Brown-Schmidt & Tanenhaus, 2006). We show that this effect extends to the timing

of speech onset. When a speaker's initial scan is more extensive, they are more sensitive to the presence of a competitor object when deciding if they have enough information to begin speaking.

Because the objects in our stimuli are so densely packed, we are unable to reliably resolve fixation points to specific objects by simply identifying the nearest object (Brown-Schmidt & Tanenhaus, 2006; Fang et al., 2009; Bock et al., 2003). Instead, we compute a holistic measurement to estimate how much visual scanning the speaker has done at a specific point in time. We call this measurement *scan distance*. To determine how much of the scene a participant has scanned, we measure the total distance their fixations have traversed, measured in object radii, at a particular time. We assume an initial fixation at the target (the location of the pre-stimulus fixation cross); the initial distance traveled is then 0 until another saccade is detected, and increases subsequently.

Scan distance measures the total distance traveled, not the distance attained from the target or any other particular point. For example, suppose a participant has viewed a 5×5 grid, with the initial fixation at the center of the target object $(2, 2)$, and subsequent saccades at .4 seconds to the top left corner, $(1, 1)$ and then at .9 seconds to the bottom right, $(5, 5)$. The scan distance at .1 seconds is 0 (since the participant has not made a saccade before this time). The scan distance at .5 seconds is 2 object radii, and the scan distance at 1 second is 10 radii (2 from the first saccade and 8 from the second). The initial dependence on the target is due only to our methodological decision to place the fixation cross on the target location.

We then conduct a post-hoc reanalysis of the onset times by adding eyetracking information to our linear model (along with the previous effects of competitor presence, grid size and scene type, and all interactions). We measure scan distance at a particular time point in order to compare the relative amounts of visual scanning in different trials. In our main analyses below, we set this time to 1 second (avoiding the uninformative early interval before saccades have occurred as well as the later period which overlaps with our participants' typical speech onsets).

However, we also conduct a sensitivity analysis to show that the observed effects are stable when scan distance is measured near 1 second, and diverge in reasonable ways for early and late measurement times. Because the distribution of scan distances is skewed right (short saccades are common, but there is a long tail of long-distance ones), we normalize by taking the logarithm. Also, because scan distance is correlated with both grid size and heterogeneity, as would be expected since larger, more complex images are harder to visually search and because this collinearity causes convergence problems, we residualize log scan distance against both grid size and heterogeneity, then z-transforming the residuals to ensure unit variance.

Figure 4 shows the raw correlation of scan distance with log onset time. In trials with no competitor (red line), scan distance has little effect on onset time. But when a competitor is present (blue line), larger

Table 5: Effects from linear mixed-effects model of log onset time with scan distance[a]

Effect	β		std. error	t value
Intercept	0.60	$***$ •	0.08	7.27
Scan distance at 1 second (resid., z-trans.)	0.04		0.03	1.46
Competitor present	0.12	$***$ •	0.03	4.50
Grid size (z-trans.)	0.04	$***$ •	0.01	3.60
Multi⇒uniform	-0.02		0.03	-0.90
Skew⇒multi	0.00		0.04	-0.13
Scan distance × competitor	0.04	$*$	0.02	2.69
Scan distance × grid size	0.00		0.01	0.23
Scan distance × multi⇒uniform	0.00		0.03	-0.13
Scan distance × skew⇒multi	-0.03		0.03	-1.00
Competitor × grid size	0.00		0.01	-0.29
Competitor × multi⇒uniform	0.07	$*$	0.03	2.40
Competitor × skew⇒multi	0.03		0.05	0.71
Grid size × multi⇒uniform	-0.01		0.03	-0.37
Grid size × skew⇒multi	0.02		0.03	0.67
Scan distance × competitor × grid size	0.02		0.01	1.27
Scan distance × competitor × multi⇒uniform	-0.07	†	0.04	-1.72
Scan distance × competitor × skew⇒multi	0.02		0.03	0.49
Scan distance × grid size × multi⇒uniform	0.03		0.04	0.79
Scan distance × grid size × skew⇒multi	0.01		0.03	0.33
Competitor × grid size × multi⇒uniform	0.05	†	0.03	1.82
Competitor × grid size × skew⇒multi	0.03		0.04	0.82
Scan distance × competitor × grid size × multi⇒uniform	-0.05		0.04	-1.34
Scan distance × competitor × grid size × skew⇒multi	0.05		0.03	1.03

[a] †: $p < .1$, $*$: $p < .05$, $**$: $p < .01$, $***$: $p < .001$; •: part of set R with expected $FDR(R) < .05$

scan distance correlates with a substantial onset delay. We interpret this as follows: The purely visual effort of searching the scene for the first competitor has little effect on onset time. However, in scenes with a competitor, onset time reflects the increased visual and linguistic effort necessary to construct a more complex description.

Model results are shown in Table 5. Confirming the pattern in the plots, the effect of scan distance on onset time differs by competitor presence (no main effect of scan distance but a competitor × scan distance interaction, $\beta = .04$). It is necessary to point out that the effect is not selected when controlling FDR at the 0.05 level, and thus should be treated with a degree of caution. But in any case, statistical controls for false positives cannot be fully reliable for post-hoc tests. While we discuss the implications of the effect below, we acknowledge that a fully convincing demonstration that it exists would require a replication of the present study with different participants.

When a speaker saccades across more of the scene in the first second, they can find a competitor early,

Figure 4: Log onset time as a function of scan distance viewed by 1 second.

and then spend a second or so (before mean speech onset at 2.1 sec) planning how best to disambiguate the target. When the initial scan does not cover much of the scene by 1 second, the speaker may find the competitor later on. Rather than revise their plan before speaking, however, they launch into the description and add information incrementally (see Figure 4; onset times for lowest scan distances are similar regardless of competitor presence). Other coefficients remain essentially unchanged from the original model, and the pattern of significance is unchanged (Table 4). None of the additional scan-distance interactions yields significance.

Sensitivity analysis

The choice of 1 second as the measurement point for scan distance is arbitrary, and a possible criticism of our analysis is that the results might depend in some crucial way on this decision. Here, we show that they do not. Figure 5 shows versions of figure 4 in which scan distance is measured at each 0.1 second time interval between .5 seconds and 1.9 seconds. In each case, the lines for competitor absent (red) and present (blue) are initially close together, but diverge as scan distance increases, with the line for competitor presence climbing while the other line stays relatively flat.

We also reran versions of the regression analysis in Table 5 with these modified scan distance measurements. (All preprocessing and coding of fixed effects remains the same, except for the initial computation of scan distance.) We re-test the significance of the interaction between scan distance and competitor presence. For the earlier sampling times (< 1), the interaction is non-significant, but at all later points, it is significant.

The primary conclusion of this analysis is that our result could be obtained by selecting any point

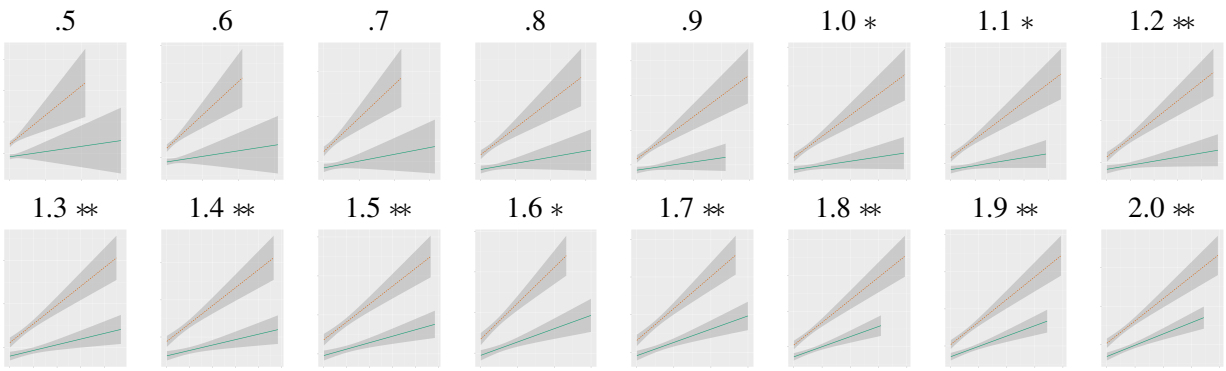

Figure 5: Log onset time (y-axis) as a function of scan distance (x-axis) at times from .5 to 1.9 seconds (compare figure 4). Significance marks refer to the interaction between competitor presence and scan distance: †: $p < .1$, ∗: $p < .05$, ∗∗: $p < .01$, ∗∗∗: $p < .001$.

from 1 to 2 seconds to measure scan distance. This is a substantial interval considering that the mean onset of speech is at 2.1 seconds.

A secondary issue is the interpretation of the changes across the different measurement times depicted in Figure 5. Early measurement times do not capture a sufficient number of saccades to provide reliable generalizations about how a participant is scanning the scene. In our dataset, the mean *latency* (occurrence time) of the first saccade after the start of the trial is 0.54 seconds. Since scan distance is 0 until the first saccade, these early scan distance measurements are uninformative for most of the trials.[12] Late measurements of scanning distance differentiate better, and also show another effect. At about 1.5 seconds, the lines for scenes with unique targets begin to trend upward, rather than remaining flat. This suggests that, after this amount of time, speakers who continue to scan a competitor-less scene energetically may be engaged in behavior associated with onset delays, for example, counting objects to produce a *coordinate* description.

6 Discussion

The analyses presented here test how visual features of a scene influence speakers' REG. We replicate previous findings that larger scenes impose delays on REG speech onset (Gatt et al., 2016), that ambiguous targets require additional disambiguation (Dale & Reiter, 1995) and that, in heterogeneous scenes, descriptions using landmarks are an important strategy for reference (Viethen & Dale, 2011; Clarke, Elsner, & Rohde, 2013). In addition to these findings, we report two novel effects. First, visually heterogeneous scenes are shown to *speed up* REG speech onset (or equivalently, visually homogenous scenes

[12] A post-hoc analysis of the time of the first saccade shows no reliable effects of scene size, type or competitor presence; the latency of the first saccade appears to reflect general rather than scene-specific processing.

delay it) in certain contexts. These heterogeneity results are in keeping with an approach in which visual properties of the scene play an important, and not necessarily interfering, role. Second, visual properties of the scene not only have an impact on REG, but they also interact with the way the speaker visually processes a particular scene: In trials where the speaker's initial scan covers little of the scene, speech onset time shows no evidence of sensitivity to the presence of a competitor.

6.1 Scene complexity and REG

We first note that our findings regarding the effect of grid size on REG go beyond similar findings from Gatt et al. (2016) due to an improvement in our methodology. In our study, participants were cued with the location of the target before the full scene appeared. This means that they did not have to visually search the scene to find their target, as they did in Gatt et al.'s study where the target was indicated with a black box around it. A concern with Gatt et al.'s method is that participants' speech onset times may reflect not only the looking and planning time, but also the time needed to initially *find* the boxed target in the scene. As such, it is possible that their observed effects of grid size reflect the participants' visual search for the target; this differs from real-world scenarios in which a speaker knows what object they want to indicate and only needs to work out how to describe it. By cuing our participants with the target location, the measured onset to speech times can be taken to reflect the time needed to scan the scene in service of utterance planning, not target discovery.

The history of research on REG focuses on the incremental nature of speakers' productions. The complexity of our scenes permits a variety of speaker strategies, and from this, we can see how the selection of an REG strategy can differ from the actual assigning of specific content or values to the description and how speakers may update their strategy as they notice properties of the scene. Our hetereogeneity results suggests that, in complex scenes, the increased difficulty of visual search can be outweighed by the linguistic helpfulness of non-attribute-based strategies like relative descriptions. Why are strategies like relative description so useful? If it is the case that a speaker can guess early in their scene processing how easy it will be to find an appropriate landmark, and how well it will uniquely identify the target, without actually finding one, they can use that information to dictate what strategy to select. Thus, the effort of planning and checking a relative description is invested only in cases where it is likely to be fast and informative.

Similarly, Bock et al. (2003) show in a time-telling task that the initial fixation can extract a coarse representation of the entire scene which guides further utterance planning. REG research finds that scene statistics guide the inclusion of color (Sedivy, 2003b; Koolen, Krahmer, & Swerts, 2015), and argues that preference among different descriptive strategies reflects adaptation to the statistics of the scene,

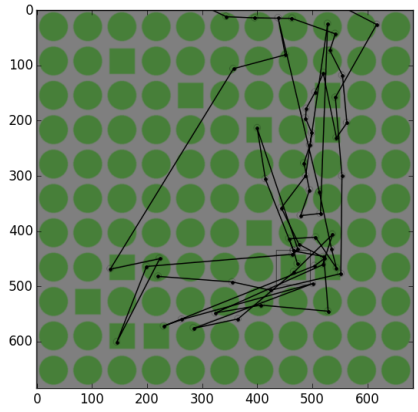

Figure 6: Scanpath corresponding to the coordinate description "ooh green square uh f[rom] on the eighth row and ninth column".

rather than a hardwired preference order (Mitchell et al., 2013b; Koolen, Krahmer, & Theune, 2012). The extension of these findings to more complex strategies reinforces this conclusion.

The dissociation of predicting the usefulness of a strategy and actually computing its value is particularly clear in the case of *coordinates*. Coordinate descriptions are available for any grid-structured stimuli (all the stimuli in our experiment) and always uniquely identify the target. In experiments on collaborative maze navigation, Garrod and Doherty (1994) find that dyads tend to converge on coordinates as a descriptive strategy, probably because, unlike landmarks, they are equally applicable to any point in the grid. But coordinates also have a substantial visual cost, since the speaker must count out objects from the edge of the image. In many cases, this involves explicitly fixating each one: our corpus contains clear examples of this behavior (figure 6). So although speakers know that they can always use coordinates, they also know that actually doing so will be time-consuming. The association of *coordinates* with uniform rather than skewed scenes suggests that, in these contexts, speakers indeed use *coordinates* as a fallback strategy when other descriptive strategies do not look promising. Again, this shows their ability to flexibly update their preference among strategies based on properties of the image.

While many REG models incorporate a flexible preference order among properties (Koolen et al., 2012), typically these preferences indicate the perceived discriminative power of the different attributes being considered (Paraboni et al., 2015). Our results imply that the speaker's selection criteria for descriptive strategies should also involve an estimate of the visual work needed to compute them. (We suspect that it also involves an estimate of linguistic difficulty— for instance, that less "nameable" objects are dispreferred as landmarks— but leave this for future work.)

At a higher level, it is clear that human REG adapts well to complex scenes. Rather than progressively

degrading as visual clutter increases, REG planning reacts by selecting strategies which avoid expensive visual searches. All the scene-dependent strategies we investigated have this property to some extent: a strategy that uses landmarks in relative descriptions exploits the "pop-out" effect; naming coordinates has a roughly constant cost regardless of the overall scene type; the inclusion of region phrases can focus the search for competitors on a small subpart of the image. These REG strategies do not replace the inclusion of simple object attributes like color; REs are highly likely to include such attributes, even redundant ones (Viethen et al., 2012; Koolen et al., 2015). But the strategies observed here act as a "force multiplier" by narrowing the search domain in which these attributes might be located.

6.2 Scanning behavior

Speakers' tendency to avoid costly visual searches can easily lead to their missing important scene information, or to discovering it late in the planning process. As discussed above, slow visual scanning has been advanced as an explanation for failure to place linguistic focus on contrastive adjectives (Pechmann, 1989) and for the production of non-canonical or disfluent adjective ordering (Pechmann, 1989; Brown-Schmidt & Tanenhaus, 2006). Our finding, that scanning too little of the scene early on also eliminates the onset time delay caused by competitor presence, fits with these findings in establishing that incrementality can lead speakers astray.

When a speaker fails to invest enough visual effort to find a competitor, they begin to speak too early, suggesting that they are going ahead with a simple initial plan which lacks sufficient disambiguating information. This result qualifies our conclusion above, that REG planning reflects effective use of scene information— if the crucial information is difficult to obtain, planning will simply carry on without it. In the framework of rational referring expression generation (Frank & Goodman, 2012; Degen & Franke, 2012; Vogel, Potts, & Jurafsky, 2013), this is a "bounded rationality" effect (Simon, 1972). The bound is imposed by the visual system's inability to supply a needed piece of information quickly or cheaply enough.

An important question for future work is what causes the variability in scanning behavior across trials and participants. Are the effects mainly due to demographic differences between participants (for instance, in their visual acuity, their alertness or their willingness to compromise on the quality of their REs)? Or are they effects of items (for instance, the placement of competitor objects or the arrangement of visually salient objects in the skewed condition)? We do find differences among participants in the distribution of scan distances, with variations in both means and standard deviations. But since there are relatively few participants to analyze, and each one saw a different set of images, it is not clear that these differences are meaningful.

We might conjecture that participants who scan the scene more might be more effective at formulating good REs, while participants who scan less might be more likely to be underspecific, to add unhelpful information, or to add helpful information in a disfluent way. But it is also possible that under some circumstances, a small scan distance indicates that the stimulus is somehow easier than normal (for instance, if the target is very close to a corner, so that coordinates are quick to obtain). In this case, the speaker may quickly devise a good RE without much effort, even if the condition might otherwise be difficult (a large, heterogeneous grid). Another possibility is that some participants spend more time checking already-formulated plans. In this case, they would fixate more (especially after onset), but not necessarily add or change anything about their REs, a point also raised in Bock et al. (2003). Such checking fixations would presumably occur more often after speech onset; we leave questions about the purpose of late fixations for future work.

Again, what this speculation underscores is the impact of the idiosyncratic nature of scenes, which speakers can and do exploit in REG planning. Observing speakers' responses to such scenes requires experiments that use sufficiently complex scenes that can elicit speakers' larger repertoire of strategies.

6.3 Implications for computational models

The results of this study imply that existing models of referring expression generation, such as the Incremental Algorithm, must be modified in order to match human behavior in timing and realization of REs as well as their content. The Incremental Algorithm and most of its descendants add one piece of content to the expression at a time. Each new piece of content triggers a new visual search to calculate the adequacy of the resulting description, and REG ends once an adequate description has been produced.

Instead, we propose that the utility[13] of different descriptive strategies are evaluated early in REG. For complex strategies, the evaluation of utility (whether the strategy would be a good choice) is potentially independent of the actual content planning— for instance, a speaker may perceive that good landmarks are available in a scene without actually selecting one. The decision of when to speak, and when to stop speaking, relies on visual search, but a full search to verify that the description is adequate may never be fully conducted. Instead, speakers balance the costs of over- or under-specification against their desire to avoid an awkward pause (Fox Tree, 2002; Sacks, Schegloff, & Jefferson, 1974) or be unnecessarily verbose.

These claims are motivated by our experimental results. Onset times remain low in complex conditions: Speakers effectively choose among the available strategies depending on scene types, implying that strategy choice is rapid and does not require a full-scale visual search for each strategy considered.

[13] In the sense of the rational modeling literature, where "utility" of an utterance for a speaker is typically defined to balance the pressure to be understood with some form of least-effort principle.

Rapid speech onset in trials with short scan distances also shows that speakers may start speaking based on a high initial utility estimate for a too-simple strategy.

A few REG models use preliminary scans which can rapidly incorporate whole-scene features in their decisionmaking. The Visible Objects Algorithm (Mitchell et al., 2013b) proposes a two-phase version of the Incremental Algorithm; it selects certain properties (color, shape, size and location) which can be quickly accessed by the visual system before conducting its object-by-object scan. This two-phase pattern was intended to explain the frequent use of these properties in descriptions even when they are redundant (Viethen et al., 2012). The attentional captioning model (Xu et al., 2015) uses neural nets to infer an initial high-level "snapshot" of the scene followed by successive areas of attention within it. While neither model attempts to directly predict speech onset timing or the presence of repairs or corrections in spoken utterances, both models permit a visual processing phase to contribute to the evaluation of multiple descriptive strategies in concert

In the vision literature, the task control model (Hayhoe & Ballard, 2014) proposes a similarly multi-modular architecture. For instance, a driving task has sub-modules for following a target car, staying on the road and avoiding obstacles, along with an executive control system which allocates visual attention between them. In an REG model, the sub-modules might correspond to descriptive strategies, with the executive system switching attention between them based on an emerging content plan.

Thus, our recommendations would move computational REG in the direction of the psycholinguistic mainstream: towards incremental models with parallel processing, controlled by a central executive which allocates limited supplies of working memory and attention. We hope future work will highlight more details of the planning process. These include a more precise picture of how fixations on individual objects affect content planning, and an improved understanding of the tradeoffs between purely linguistic factors like "nameability" and visual costs.

7 Conclusion

Our results are in keeping with other work in speech production that demonstrates the non-encapsulation of different processing stages: Whereas that prior work (e.g., Vigliocco & Hartsuiker, 2002) focuses on the contributions of sound, syntax, and meaning in sentence production, our work highlights the interdependence between speakers' linguistic decisions about how to convey and order information and their visual processing for accessing that information.

- we're looking at the ⟨relative⟩only blue square⟨/relative⟩ ⟨region⟩on the screen⟨/region⟩ actually it's the ⟨relative⟩only square⟨/relative⟩ ⟨region⟩on the screen⟨/region⟩ it's ⟨region⟩towards the centre⟨/region⟩ the top centre ⟨region⟩of the screen⟨/region⟩
- the ⟨target⟩darker red circle⟨/target⟩ in the ⟨coord⟩fourth row⟨/coord⟩ from ⟨region⟩the top⟨/region⟩ and the ⟨coord⟩second object⟨/coord⟩ from the right ⟨lmark⟩next to the big bright red square⟨/lmark⟩
- the ⟨target⟩blue square⟨target⟩
- ⟨target⟩green circle⟨/target⟩ the ⟨relative⟩only green circle⟨/relative⟩ towards the ⟨lmark⟩left of a bright big square⟨/lmark⟩

Figure 7: Some sample tagged utterances.

8 Acknowledgements

This research was supported by EPSRC grant EP/H050442/1 and European Research Council grant 203427 "Synchronous Linguistic and Visual Processing". We thank Marten van Schijndel for discussion of mixed effect models, research assistant Emma Ward for help in transcription and Amelia Hunt for her comments on a draft of the manuscript. We also thank the helpful comments and suggestions from three anonymous reviewers.

9 Appendix: Deterministic phrase tagger

We begin by grouping the lexical items used into equivalence classes such as *color* ("red", "blue"...) and *region* ("top", "quadrant", "south"...). There are 147 word types in our lexicon; the remaining items (such as determiners and conjunctions) are classed as *other*. We then apply a series of regular expressions, one at a time, in order. These are the following:

1. Tag *relative (color|size)* (shape|one)* as "Relative" (*only green circle*)
2. Tag *cardinal? (color|size)* (shape|one)* as "Object" (*two green circles*)
3. Tag *prep other* object+* as "Landmark" (*beside the OBJ*)
4. Tag *prep other* region+* as "Region" (*in the corner*)
5. Tag *(cardinal|ordinal|one) (row|col)? (other|prep)* region** as "Coordinate" (*two rows down from the top*)
6. Tag *prep* its prep* as "Landmark" (*on its right*)
7. Tag *(prep|other)* prep* as "Region" (*on the right*)
8. Tag all "Object" and remaining size and color descriptors as "Target"

10 Appendix: Descriptors by condition

Table 6, below, gives the percentage of descriptions including *color*, *shape* and *size* terms by condition. Note that proportions are roughly uniform across conditions except for *size*, which is used primarily in the skewed condition, the only one where objects vary in size.

Condition	Color	Shape	Size	N
Uniform	27%	33%	2%	2028
Multicolor	30%	33%	3%	1956
Skewed	30%	33%	8%	2154

Table 6: Percentage of descriptions including color, shape and size descriptors, by condition.

References

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of memory and language*, *38*(4), 419–439.

Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science*, *12*(2), 157–162.

Arnold, J., Eisenband, J., Brown-Schmidt, S., & Trueswell, J. (2000). The rapid use of gender information: Evidence of the time course of pronoun resolution from eyetracking. *Cognition*, *76*(1), B13-B26.

Asher, M. F., Tolhurst, D. J., Troscianko, T., & Gilchrist, I. D. (2013). Regional effects of clutter on human target detection performance.

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*(3), 255–278.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects models using eigen and s4 [Computer software manual]. Available from `http://CRAN.R-project.org/package=lme4` (R package version 1.1-7)

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, 289–300.

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, 1165–1188.

Bock, K., Irwin, D. E., Davidson, D. J., & Levelt, W. (2003). Minding the clock. *Journal of Memory and Language*, *48*(4), 653 - 685.

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

Clark, H. H., & Tree, J. E. F. (2002). Using uh and um in spontaneous speaking. *Cognition*, *84*(1), 73–111.

Clarke, A., Coco, M. I., & Keller, F. (2013). The impact of attentional, linguistic, and visual features during object naming. *Frontiers in Perception Science, Special Issue on Scene Understanding*, *4(329)*.

Clarke, A., Elsner, M., & Rohde, H. (2013). Where's Wally: the influence of visual salience on referring expression generation. *Frontiers in Perception Science, Special Issue on Scene Understanding*, *4(329)*, 1-10.

Clarke, A., Elsner, M., & Rohde, H. (2015). Giving good directions: order of mention reflects visual salience. *Frontiers in psychology*, *6*.

Clarke, A., Green, P., Chantler, M., & Emrith, K. (2008). Visual search for a target against a 1/fβ continuous textured background. *Vision research*, *48*(21), 2193–2203.

Clarke, A., Green, P., Chantler, M., & Hunt, A. (2016). Human search for a target on a textured background is consistent with a stochastic model. *Journal of Vision*.

Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology*, *6*(1), 84–107.

Curry, A. C., Gkatzia, D., & Rieser, V. (2015). Generating and evaluating landmark-based navigation instructions in virtual environments. *ENLG 2015*, 90.

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

Degen, J., & Franke, M. (2012). Optimal reasoning about referential expressions. In *Proceedings of SEMDial.*

Dell, G. S. (1986). A spreading activation theory of retrieval in language production. *Psychological Review*, *93*, 283-321.

Duan, M., Elsner, M., & Marneffe, M.-C. de. (2013). Visual and linguistic predictors for the definiteness of referring expressions. In *Proceedings of the 17th workshop on the semantics and pragmatics of dialogue (SemDial).* Amsterdam.

Eckstein, M. P. (2011). Visual search: A retrospective. *Journal of Vision*, *11*(5).

Ehinger, K. A., Hidalgo-Sotelo, B., Torralba, A., & Oliva, A. (2009). Modelling search for people in 900 scenes: A combined source model of eye guidance. *Visual cognition*, *17*(6-7), 945–978.

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

Fang, R., Chai, J. Y., & Ferreira, F. (2009). Between linguistic attention and gaze fixations in multimodal conversational interfaces. In *Proceedings of the 2009 international conference on multimodal interfaces* (pp. 143–150). New York, NY, USA: ACM. Available from `http://doi.acm.org/10.1145/1647314.1647339`

Ferreira, F., & Swets, B. (2002). How incremental is language production? evidence from the production of utterances requiring the computation of arithmetic sums. *Journal of Memory and Language*, *46*, 57-84.

Ferreira, V. S. (1996). Is it better to give than to donate? syntactic flexibility in language production. *Journal of Memory and Language*, *35*, 724-755.

Ford, W., & Olson, D. (1975). The elaboration of the noun phrase in children's description of objects. *Journal of Experimental Child Psychology*, *19*, 371-382.

Fox Tree, J. E. (2002). Interpreting pauses and ums at turn exchanges. *Discourse Processes*, *34*(1), 37–55.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998–998.

Fukumura, K., van Gompel, R. P., Harley, T., & Pickering, M. J. (2011). How does similarity-based interference affect the choice of referring expression? *Journal of Memory and Language*, *65*(3), 331 - 344. Available from `http://www.sciencedirect.com/science/article/pii/S0749596X1100060X`

Garrod, S., & Doherty, G. (1994). Conversation, co-ordination and convention: An empirical investigation of how groups establish linguistic conventions. *Cognition*, *53*(3), 181–215.

Gatt, A., Krahmer, E., Deemter, K. van, & van Gompel, R. P. G. (2016). Reference production as search: The impact of domain size on the production of distinguishing descriptions. *Cognitive Science*.

Gatt, A., van Gompel, R. P. G., Krahmer, E., & van Deemter, K. (2012). Does domain size impact speech onset time during reference production? In *Proceedings of the 34th Annual Meeting of the Cognitive Science Society* (p. 1584-1589). Sapporo, Japan: Cognitive Science Society.

Gilchrist, I. D., & Harvey, M. (2006). Evidence for a systematic component within scan paths in visual search. *Visual Cognition*, *14*(4-8), 704–715.

Griffin, Z. M., & Bock, K. (2000). What the eyes say about speaking. *Psychological science*, *11*(4), 274–279.

Hayhoe, M., & Ballard, D. (2014). Modeling task control of eye movements. *Current Biology*, *24*(13), R622–R628.

Klippel, A., & Winter, S. (2005). Structural salience of landmarks for route directions. In A. Cohn & D. Mark (Eds.), *Spatial information theory, international conference cosit* (p. 347-362). Berlin: Springer.

Koolen, R., Gatt, A., Goudbeek, M., & Krahmer, E. (2011). Factors causing overspecification in definite descriptions. *Journal of Pragmatics*, *43*(13), 3231 - 3250. Available from `http://www.sciencedirect.com/science/article/pii/S0378216611001731`

Koolen, R., Krahmer, E., & Swerts, M. (2015). How distractor objects trigger referential overspecification: Testing the effects of visual clutter and distractor distance. *Cognitive Science*.

Koolen, R., Krahmer, E., & Theune, M. (2012). Learning preferences for referring expression generation: Effects of domain, language and algorithm. In *Proceedings of the seventh international natural language generation conference* (pp. 3–11). Stroudsburg, PA, USA: Association for Computational Linguistics. Available from `http://dl.acm.org/citation.cfm?id=2392712.2392718`

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

Landy, M. S., & Bergen, J. R. (1991). Texture segregation and orientation gradient. *Vision research*, *31*(4), 679–691.

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

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*, 1-75.

Meyer, A. S., Sleiderink, A. M., & Levelt, W. J. (1998). Viewing and naming objects: eye movements during noun phrase production. *Cognition*, *66*(2), B25 - B33.

Mitchell, M., van Deemter, K., & Reiter, E. (2013a). Attributes in visual reference. In *Proceedings of the Production of Referring Expressions workshop at the conference of the Annual Meeting of the Cognitive Science Society.*

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

Paraboni, I., Galindo, M. R., & Iacovelli, D. (2015). Generating overspecified referring expressions: the role of discrimination. In *53rd annual meeting of the association for computational linguistics (acl-2015)* (pp. 76–82). Beijing: Association for Computational Linguistics.

Pechmann, T. (1989). Incremental speech production and referential overspecification. *Linguistics*, *27*(1), 3 - 1756.

Pyykknen, P., & Jrvikivi, J. (2010). Activation and persistence of implicit causality information in spoken language comprehension. *Experimental Psychology*, *57*, 5-16.

Reali, F., Spivey, M. J., Tyler, M. J., & Terranova, J. (2006). Inefficient conjunction search made efficient by concurrent spoken delivery of target identity. *Perception and Psychophysics*, *68*, 959–974.

Richter, K.-F. (2008). *Context-specific route directions: Generation of cognitively motivated wayfinding instructions.* DisKi 314 / SFB/TR 8 Monographs Volume 3. Amsterdam: IOS

Press.

Rosenholtz, R., Li, Y., & Nakano, L. (2007). Measuring visual clutter. *Journal of Vision*, *7*, 1-21.

Rubio-Fernández, P. (2016). How redundant are redundant color adjectives? an efficiency-based analysis of color overspecification. *Frontiers in Psychology*, *7*(153).

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *language*, 696–735.

Sedivy, J. C. (2003a). Pragmatic versus form-based accounts of referential contrast: Evidence for effects of informativity expectations. *Journal of Psycholinguistic Research*, *32*(1), 3-23. Available from `http://dx.doi.org/10.1023/A%3A1021928914454`

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

Shriberg, E. E. (1994). *Preliminaries to a theory of speech disfluencies*. Unpublished doctoral dissertation, University of California at Berkeley.

Simon, H. A. (1972). Theories of bounded rationality. *Decision and organization*, *1*(1), 161–176.

Sonnenschein, S. (1985). The development of referential communication skills: Some situations in which speakers give redundant messages. *Journal of Psycholinguistic Research*, *14*(5), 489–508.

Spain, M., & Perona, P. (2010). Measuring and predicting object importance. *International Journal of Computer Vision*, *91*, 59-76.

Spivey, M. J., Tyler, M. J., Eberhard, K. M., & Tanenhaus, M. K. (2001). Linguistically mediated visual search. *Psychological Science*, *12*, 282–286.

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

Tatler, B. W., Hayhoe, M. M., Land, M. F., & Ballard, D. H. (2011). Eye guidance in natural vision: Reinterpreting salience. *Journal of Vision*, *11*(5).

Treisman, A. (1985). Preattentive processing in vision. *Computer vision, graphics, and image processing*, *31*(2), 156–177.

Viethen, J., & Dale, R. (2011). GRE3D7: A corpus of distinguishing descriptions for objects in visual scenes. In *Proceedings of the workshop on using corpora in natural language generation and evaluation.* Edinburgh, Scotland: Association for Computational Linguistics.

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 2012).* Sapporo, Japan: Cognitive Science Society.

Vigliocco, G., & Hartsuiker, R. J. (2002). The interplay of meaning, sound, and syntax in sentence production. *Psychological Bulletin*, *128*, 442-472.

Vogel, A., Potts, C., & Jurafsky, D. (2013). Implicatures and nested beliefs in approximate decentralized-pomdps. In *Acl (2)* (pp. 74–80).

Waller, D., & Lippa, Y. (2007). Landmarks as beacons and associative cues: Their role in route learning. *Memory and Cognition*, *35*.

Whitehurst, G. J. (1976). The development of communication: Changes with age and modeling. *Child Development*, *47*(2), 473-482.

Wolfe, J. M. (2012). Visual search. In P. Todd, T. Holls, & T. Robbins (Eds.), *Cognitive search: Evolution, algorithms and the brain* (p. 159 - 175). Cambridge, MA, USA: MIT Press.

Wolfe, J. M., Võ, M. L.-H., Evans, K. K., & Greene, M. R. (2011). Visual search in scenes involves selective and nonselective pathways. *Trends in cognitive sciences*, *15*(2), 77–84.

Xu, K., Ba, J. L., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., et al. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of icml.* Lille, France: JMLR.

Yuan, J., & Liberman, M. (2008). Speaker identification on the SCOTUS corpus. *Journal of the Acoustical Society of America*, *123*(5), 3878.