Integrating Mechanisms of Visual Guidance in Naturalistic Language Production

Moreno I. Coco and Frank Keller
Institute for Language, Cognition and Computation
School of Informatics, University of Edinburgh
10 Crichton Street, Edinburgh EH8 9AB, UK
Phone: +44 131 650 8289, Fax: +44 131 650 4587
mcoco@staffmail.ed.ac.uk, keller@inf.ed.ac.uk

Abstract

Situated language production requires the integration of visual attention and linguistic processing. Previous work has not conclusively disentangled the role of perceptual scene information and structural sentence information in guiding visual attention. In this paper, we present an eye-tracking study that demonstrates that three types of guidance, perceptual, conceptual, and structural, interact to control visual attention. In a cued language production experiment, we manipulate perceptual (scene clutter) and conceptual guidance (cue animacy), and measure structural guidance (syntactic complexity of the utterance). Analysis of the time course of language production, before and during speech, reveals that all three forms of guidance affect the complexity of visual responses, quantified in terms of the entropy of attentional landscapes and the turbulence of scan patterns, especially during speech. We find that perceptual and conceptual guidance mediate the distribution of attention in the scene, whereas structural guidance closely relates to scan-pattern complexity. Furthermore, the eye-voice span of the cued object and its perceptual competitor are similar; its latency mediated by both perceptual and structural guidance. These results rule out a strict interpretation of structural guidance as the single dominant form of visual guidance in situated language production. Rather, the phase of the task and the associated demands of cross-modal cognitive processing determine the mechanisms that guide attention.

Keywords: Eye-movements; language production; scene understanding; cross-modal processing; eye-voice span; structural guidance.

Introduction

In tasks where language is produced in a visual context, such as giving directions on a map, explaining the function of a device, or telling the time, visual and linguistic information have to be synchronized. Such synchronization draws upon cross-modal cognitive mechanisms which allow

different modalities to share, exchange, and integrate information. Psycholinguistic research on language production in a visual context has demonstrated that factors from both visual processing and sentence processing can influence language production.

On one hand, there is evidence that perceptual factors are involved in selecting referential information for sentence processing (e.g., Arnold & Griffin, 2007; Gleitman, January, Nappa, & Trueswell, 2007; Papafragou, Hulbert, & Trueswell, 2008; see also Myachykov, Thompson, Scheepers, & Garrod, 2011, for a review). On the other hand, there is also evidence for sentence processing mechanisms which explain how eye-movements and speech responses are linked (Meyer, Sleiderink, & Levelt, 1998; Griffin & Bock, 2000; Bock, Irwin, Davidson, & Levelt, 2003; Qu & Chai, 2008; Kuchinksy, Bock, & Irwin, 2011). The *eye-voice span* (EVS) is the most well-studied manifestation of such mechanisms: visual objects are looked at approximately 900 ms before being mentioned, providing evidence for how visual attention is *structurally* guided by sentence processing.

Previous research has uncovered a range of factors, linguistic and non-linguistic, that are involved in situated language processing. Visual attention uses scene information employing search routines guided by the available perceptual material, while sentence information drives visual search towards linguistic targets selected for production. Each message that is planned entails a precise set of referent objects; these are visually attended according to the structure of the sentence being produced. A realistic account of situated language production must explain how these different sources of information interact when guiding visual attention.

Crucial to this understanding are results by Kuchinksy et al. (2011), which follow up on Bock et al.'s (2003) work, and try to establish whether visual attention is guided by perceptual resources related to the task, by structural regularities of the utterance produced, or by an interaction between the two. In Kuchinksy et al.'s (2011) study, participants were asked to read aloud the time, e.g., *half past three* from drawings of analogue clocks. The two types of guidance, structural and perceptual, were tested by manipulating the length of the hands of the clock and the instructions for the task. In one condition, the hour was indicated by a short hand and the minutes by a long hand (perceptually typical), in the other condition, this was reversed: long hand for hour, short hand for minutes (perceptually atypical). Moreover, participants were instructed to read normally (short hand for hours, long hand for minutes) or inverted. The results show that participants read the time according to the instructions given, regardless of whether the hands match their perceptual experience of clocks, i.e., hours are usually indicated by a short hand. In light of these findings, the authors suggest that perceptual and structural guidance do not coexist; rather, structural guidance plays a leading role in directing visual attention during situated production, and the two types of guidance do not seem to act together.

However, Kuchinsky et al.'s view is problematic in a number of ways, starting with their definition of structural vs. perceptual guidance, which is highly dependent on their particular task of telling the time. Here, we advocate the view that structural guidance should refer to mechanisms that are germane to sentence processing. A more appropriate definition would be one that refers to quantifiable linguistic properties of a sentence, and that can therefore be used to determine how the linguistic components of structural guidance (e.g., types and frequencies of particular syntactic phrases) affect visual responses. In a naturalistic context, structural guidance should be related to structural information associated with the sentence produced (e.g., its syntactic structure). The approach we advocate should make it easier to disentangle task-dependent factors from purely structural guidance factors.

Similarly, although perceptual guidance is generally taken to refer to low-level properties of the visual information (Gleitman et al., 2007), such definition is not always consistent across studies. For example, Kuchinksy et al. (2011) use the term to refer to experience of reading clocks, which again reflects task-based, conceptual knowledge (short hand for hours, long hand for minutes). Thus, to systematically investigate perceptual guidance independent of task, it is necessary to use visual stimuli which allow the explicit manipulations of low-level visual features, as opposed to visually impoverished stimuli that are likely to underestimate the importance of perceptual effects.

To address these concerns, this article presents an experimental framework that provides a more direct evaluation of different types of guidance in visually situated language production. In our view, perceptual and structural guidance in situated language production are not mutually exclusive, and not the only possible types of guidance.

There is at least another form of guidance, which emerges from the *conceptual* properties of the referents attended. Compared to inanimate referents, animates are mentioned more frequently, and are more likely to take the role of sentential subject and to be pronominalized (e.g., Fukumura & Van Gompel, 2011). Moreover, animacy plays a crucial role on visual attention: animate objects are localized faster and fixated for longer than inanimate objects (Fletcher-Watson, Findlay, Leekam, & Benson, 2008). These results suggest that the conceptual properties of an object, such as its animacy, have important repercussions on the allocation of visual attention in the scene, as well as on the utterance chosen to refer to it.

This leads us to an overall theoretical framework which makes it possible to investigate and quantify the role that each type of guidance plays during situated production:

Perceptual guidance steers visual attention to interesting locations, which may carry information relevant to sentence encoding. This definition of perceptual guidance is compatible with that of visual *saliency* (Itti & Koch, 2000; Parkhursta, Lawb, & Niebur, 2002). Saliency, as a composite measure of low-level visual information, is positively correlated with the presence of objects (Elazary & Itti, 2008), and found to guide attention in tasks where targets are underspecified (e.g., memorization, Underwood & Foulsham, 2006), and also utilized during situated language production (Gleitman et al., 2007). In our view, the most appropriate way to define visual saliency is in terms of visual clutter, calculated by integrating low-level visual information, e.g., color, with edge information (Rosenholtz, Mansfield, & Jin, 2005). The inclusion of edges sets apart clutter from standard visual saliency (Itti & Koch, 2000), which is based solely on low-level features, without taking object-based visual information such as edges into account. We select visual clutter because it gives us a better estimate of objecthood than standard saliency measures, and therefore is more suitable for studies of situated language processing, where visual objects act as linguistic referents.

Conceptual guidance is a high-level process which draws upon a semantic analysis of the objects attended to compute their contextual relevance within the scene. Evidence for this type of guidance comes primarily from visual search experiments, where semantic information about visual referents is used to optimize the allocation of visual attention (e.g., Henderson & Hollingworth, 1999; Findlay & Gilchrist, 2001; Henderson, 2003; Zelinsky & Schmidt, 2009; Nuthmann & Henderson, 2010; Hwang, Wang, & Pomplun, 2011). Since a central conceptual feature of referents is their *animacy*, (i.e., whether they are living things or not), already shown to impact grammatical assignment and word ordering (e.g., McDonald, Bock, & Kelly, 1993; Levelt, Roelofs, & Meyer, 1999; Prat-Sala & Branigan, 2000; Branigan, Pickering, & Tanaka, 2008; Coco & Keller, 2009) and expected to play a key role in sentence production (McDonald et al., 1993), in the current study, we manipulate animacy to investigate conceptual guidance effects.

Structural guidance offers the third type of guidance mechanisms: it constrains visual attention based on the form a sentence is expected to take, thus directing attention to the objects the sentence is referring to. Such guidance should be more prominent when visual and linguistic processing have to cooperate overtly, e.g., during the mention of a referent – the two information streams need close synchronization in this phase. A prominent example of structural guidance is the fact that looks to a visual object occur in a predictable fashion prior to its linguistic mention (eye-voice span): for instance, when telling the time, looks at the long hand occur shortly before the minutes are mentioned (e.g., Bock et al., 2003). In the current study, we take a broader perspective on structural guidance, operationalizing it in terms of syntactic information. This allows us to examine how different syntactic constructions mediate visual responses.

In summary, our study is designed to bring together perceptual, structural and conceptual guidance, explored separately in previous research. In a unified framework, we examine how these three types of guidance interact in the allocation of visual attention at different phases of a sentence production task. Our general aim is to provide a more comprehensive understanding of situated language processing, in which a range of cross-modal control mechanisms can interact.

Experiment

Our basic assumption in this study is that visual processing mechanisms can only be studied comprehensively in naturalistic scenes, which are rich in number and type of objects and their relationships. This holds in particular for perceptual guidance, which is based on low-level visual information extracted from the whole scene, and conceptual guidance, which exploits the semantic properties of objects and their contextual relatedness. Thus, in our design, we situate language processing in a naturalistic visual context. This contrasts with research using the classic Visual World Paradigm (VWP, e.g., Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995), which typically employs simplistic visual scenes based on object arrays or clip-art scenes with a small number of objects.

Together with naturalistic visual responses, we also aim to elicit naturalistic linguistic responses. For this, we use a production paradigm in which participants generate a scene description after being cued with the name of a target object, which can be either animate or inanimate. Cuing is a form of conceptual guidance, as we steer the production process towards particular objects depicted in the scene, which can be expected to trigger strategic routines of visual search.

In order to increase the range of possible productions, we systematically introduce referential ambiguity for the cued object, by including two possible referents for the target object in a given scene. For example, two MEN and two CLIPBOARDS are depicted in our example scene.[1]

Method

In this experiment, participants had to describe photo-realistic indoor scenes (24 scenes drawn from six scenarios, e.g., Kitchen), after being prompted with a cue word that they were told to always use in their descriptions. There was no restriction on the amount of speech participants could produce. No addressee was given, and the purpose of the task was to provide an exhaustive description of the cued target. The full instructions are given in the Supplementary Material.

The design crossed the factors *Clutter* and *Cue*. The factor Clutter manipulated the amount of visual clutter in the scene (levels: Minimal and Cluttered), and the factor Cue manipulated the ani-

[1] Note we use small caps to denote VISUAL REFERENTS and italics to denote *linguistic referents*.

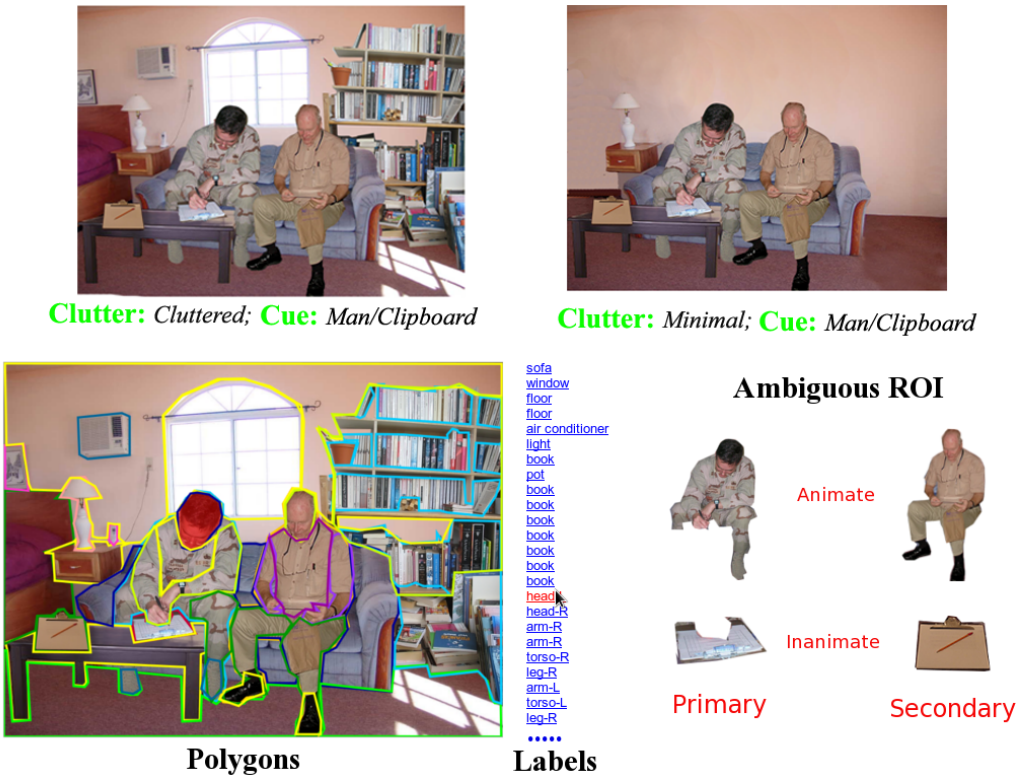


Figure 1. The top panel gives example for experimental materials in the Cluttered (left) and Minimal (right) conditions of the factor Clutter. The bottom panel (left) shows an example of same visual scene annotated with polygons and labels. The bottom panel (right) shows the ambiguous regions of interest, to which we refer as primary and secondary mention to distinguish between them.

macy of the cue word (levels: Animate and Inanimate). Each cue word ambiguously corresponded to two visual objects in the scene. Example materials are given in the top panel of Figure 1 for the cue words *man* and *clipboard*; the full set of scenes in both condition of clutter can be found in the Supplementary Material.

A basic scene containing minimal visual information was selected from existing datasets (e.g., LabelMe, Russell, Torralba, Murphy, & Freeman, 2008) or downloaded from the Internet. We used Photoshop to paste the target objects (e.g., MAN and CLIPBOARD) into the scene and added distractors to obtain the cluttered version of the same scene. A paired-samples t-test was conducted to test that the visual clutter (measured as Feature Congestion, Rosenholtz et al., 2005) in Minimal scenes (mean $= 2.78$, SD $= 1.12$) was significantly smaller than in their Cluttered version (mean $= 3.52$, SD $= 0.33$; $t(46) = 3.71$, $p = 0.0005$).

Each scene was fully annotated using the LabelMe toolbox by drawing bounding polygons around the objects in the scene and labeling them with words. These polygons were used to map the fixation coordinates onto the corresponding labels. Objects can be embedded into other objects (e.g., the head is part of the body); in this case, the smallest object that contained the fixation was used. The mean number of objects per image was 28.65 (SD $= 11.30$).

In addition to the 24 experimental items there were 48 fillers, in which the number of visual

referents corresponding to the cue varied from one to three.

Twenty-four native speakers of English, all students of the University of Edinburgh, gave informed consent and were paid five pounds for taking part in the experiment. They each saw 72 experimental items randomized and distributed in a Latin square design that made sure that each participant only saw one condition per scene.

An EyeLink II head-mounted eye-tracker was used to monitor participants' eye-movements with a sampling rate of 500 Hz. Images were presented on a 21" Multiscan monitor at a resolution of 1024×768 pixels; participants' speech was recorded with a lapel microphone. Only the dominant eye was tracked, as determined by a simple parallax test. At the beginning of each trial, a written cue word appeared for 750 ms at the center of the screen, after which the scene followed and sound recording was activated. The relatively long exposure to the cue makes sure that the first few fixations are not driven by comprehension process, as understanding the word should have been already completed prior scene onset.

Drift correction was performed at the beginning of each trial. Four practice trials were administered at the beginning of the experiment to familiarize participants with the task. There was no time limit on the duration of a trial; to pass to the next trial, participants pressed a button on a response pad. The experimental task was explained to participants using written instructions and took approximately 45 minutes to complete.

Research Questions

We focus on two main aspects of the sentence production task: the interaction between visual guidance mechanisms, and referential ambiguity resolution.

1. **How do the different types of guidance interact?** We focus on two main temporal phases of the task: before the onset of speech and during speaking. By adopting this division, we do not make any claims regarding the distinction between apprehending an event (i.e., recognizing and representing the visual context of a message) and formulating the message (i.e., the other stages involved in the production process). Rather, we want to distinguish between visual attention before and during language production. Before speech is overtly produced, visual attention is relatively independent of sentence production mechanisms, and therefore we expect a weaker influence of structural guidance on it. During speech production, on the other hand, we expect a more pronounced influence of linguistic information in guiding visual attention. During both phases, we investigate two key dependent variables: the spatial distribution of fixations and the sequential complexity of visual responses; and determine how they are influenced by perceptual, conceptual, and structural guidance.

2. **How is referential ambiguity resolved?** This has been the most studied aspect of situated production in previous work (e.g., Griffin & Bock, 2000). Referential ambiguity is known to cause interference when producing referring expressions (Arnold & Griffin, 2007; Fukumura, Van Gompel, & Pickering, 2010); and it can be successfully resolved if detected before the production starts (Ferreira, Slevc, & Rogers, 2007). Referential competition and ambiguity resolution strategies are indexed by visual responses; here we focus on the latency of last fixation on the target object before it is mentioned. We investigate how such latency is strictly mediated by a range of factors, such as the structure of the sentence, the availability of visual information, the animacy of target, as well as oculo-motor variables emerging in a naturalistic scene context.

Experimental Measures of Cognitive Processing

The aim of this experiment is to test how different forms of guidance (perceptual, conceptual, and structural) modulate visual attention during a cued language production situated in photo-realistic scenes. We hypothesize that the different forms of guidance interact in order to synchronize the two streams of processing. This should be visible both in traditional latency measures pertaining to the access and mention of referents, and in more global measures, such as the spatial distribution of fixations, and in the complexity of visual responses, as outlined below.

Attentional Complexity. The first set of questions concerns sentence production as a process demanding the integration of different cognitive information, to obtain a general understanding of the different visual guidance mechanisms involved.

Spatial Aspects. We analyze the spatial distribution of fixations before and during speech, which reflects the spread of visual attention across the scene, and indicates access to scene information during language production. We quantify the spread of attention in terms of entropy: the more spread out fixations are, the higher the entropy (see Frank, Vul, & Johnson, 2009 for a similar use of entropy). Perceptual and conceptual guidance have been shown to impact visual search performance (e.g., Henderson, Chanceaux, & Smith, 2009; Zelinsky & Schmidt, 2009, but their role in more complex cross-modal tasks remains unexplored.

In our study, we expect that the more the visual clutter, the more spread out the fixations. The larger availability of visual information forces attention to evaluate a wider section of the scene. This is especially true during speech, where visual material has to be rapidly sourced to sustain the production process.

We also expect inanimate objects to trigger wider visual sampling than animate objects before production. Animate objects are found more quickly (Fletcher-Watson et al., 2008) and are conceptually more accessible than inanimate ones (Levelt et al., 1999), hence facilitating the process of message planning before speech starts (Coco & Keller, 2009). However, during speech, animate objects need to be contextualized with the rest of the scene, entailing a wider visual sampling than inanimate object, and hence a more spread out distribution of fixations.

Temporal Aspects. Assuming a general concept of structural guidance, which uses the constituent structure of a sentence to define its sequential complexity, we hypothesize that differences in constituent structure (e.g., types of phrases) give rise to differences in the sequential complexity of visual responses, represented as scan patterns (i.e., temporal sequences of fixated objects, Noton & Stark, 1971). Scan pattern complexity is quantified as turbulence, which is a measure of its sequential variability (refer to Appendix B for details). We aim at quantifying the structural links between visual and linguistic responses by abstracting away from their referential information, e.g., the semantic properties of the object fixated or mentioned.

Our main prediction is that the complexity of the constituent structure should correlate with scan pattern complexity, especially during speaking, when visual and linguistic processing need overt synchronization. Moreover, we expect scan pattern complexity to be differentially influenced by the type of syntactic phrases composing the sentence, and be mediated by scene clutter and target animacy.

At the finer granularity of structural phrases, we predict that noun phrases increase the complexity of scan patterns less than verb phrases: NPs have usually a one-to-one relation with a specific object (e.g., *the apple* relates to the object APPLE), whereas VPs relate to a cluster of contextually related objects (e.g., *eats an apple* relates to the objects APPLE, MOUTH, HAND).

Perceptual and conceptual mechanisms should show similar pattern of effects to those obtained with the entropy of the fixation distribution. In particular, descriptions of animate targets, and cluttered scenes would demand more complex visual responses, especially during speaking. Moreover, in this finer structural analysis, we expect these two factors to interact in a positive way. As argued before, information about animate referents is easier to access, and conceptually wider than for inanimate referents; hence increasing the range of possible descriptions, and the complexity they induce in the associated eye-movement responses.

Latency. Our second set of predictions is based on the latencies of fixation on the object selected for mention (i.e., the *referent*) relative to the onset of its linguistic realization; we compare these latencies with those to the other object of the same type (i.e., the *competitor*), again relative to the mention of the referent.

Latencies are informative about referential competition and ambiguity resolution, and the effect that different types of guidance have on them. We measure the latency of the first and last fixation prior to onset of a mentioned visual referent. In the paper, we focus only on the latter, because it returns the most significant insights about the interaction between the mechanisms of visual guidance. However, we also report results of first fixation in the Supplementary Material to allow interested readers to make a comparison between our study and other work on situated language production, which has used the latency of the first fixation as the measure of interest (e.g., Brown-Schmidt & Tanenhaus, 2006).

The last fixation prior to mention, also known as eye-voice span (*EVS*), has played a seminal role in psycholinguistic studies of reading aloud (Levin & Buckler-Addis, 1979). In the context of situated production, it is widely agreed that the last fixation on a referent target is closely aligned with its mention (Griffin & Bock, 2000; Bock et al., 2003; Qu & Chai, 2008, 2010; Kuchinksy et al., 2011). The theoretical explanation is provided by a structural model of language production where visual attention is guided by the incremental construction of phrases, i.e., a look-and-mention routine. Thus, if EVS is solely influenced by a systematic application of structural guidance, then only the referent to be mentioned should be targeted by visual attention. However, given that the visual presence of a competitor influences the production of referring expressions (Fukumura et al., 2010), we expect the competitor object to be visually attended in a similar way as the referent object.

We also expect structural guidance to mediate the EVS. At this stage in the production process, visual information contextually related to the mentioned referent gets evaluated in the light of the syntactic form of the message being formulated. In particular, given our definition of structural guidance based on syntactic structure, we expect that sentences containing more nouns should be associated with longer EVS. More referents will be produced in the sentence, the more visual objects would need to be evaluated. This would anticipate the EVS time to the referent under production, and allow attention focusing on other relevant information to be sourced on the visual scene.

Moreover, if structural guidance is the only mechanism driving EVS, as Kuchinksy et al. (2011) argue, we should not find any significant influence of perceptual or conceptual mechanisms on it, nor interactions of these mechanisms with structural guidance. If, instead, visual guidance relies also on these other mechanisms to support sentence production, then the EVS should be modulated by them. In particular, the amount of visual clutter should slow the identification of the target (Henderson et al., 2009). However, since our task is linguistic, and not simply a search task, and given that the target object expresses conceptual properties such as animacy, we predict that the effect of clutter will depend on the animacy of the mentioned target. In particular, we expect an

inanimate target to have shorter EVS in scenes with minimal clutter. Such a result would be in line with findings in visual search, and advance previous work by showing that similar mechanisms of visual guidance are operating during a language production task.

We also examine the impact of oculomotor control factors, e.g., gaze duration, on latencies. It is known from research in reading (Rayner, 1998) that oculomotor mechanisms (e.g., regressions) can play an important role in eye-movements, and co-vary with other variables of interest (e.g., lexical frequency, Kennedy & Pynte, 2005; Pynte, New, & Kennedy, 2008). Similarly, when viewing a scene, there may be several co-variates, such as the number of objects fixated, or the gaze duration of these fixations, which can influence responses. Quantification of these influences on latencies helps determine the extent of the interaction between other experimental variables under investigation such as animacy, clutter, and syntactic structure.

Data Analysis

In order to better understand the dynamics operating during the concurrent processing of visual and linguistic information, the complexity of our stimuli and responses requires the inclusion of a wide range of variables in the analysis, as well as the employment of novel analytical techniques.

Independent Variables for Investigating Guidance

To investigate perceptual guidance, we manipulate visual clutter (Rosenholtz et al., 2005), a measure of visual complexity integrating low-level features (e.g., color, intensity, and orientation) and edge information. To test the impact of conceptual guidance, we manipulate the animacy of the cue that orients the speaker towards a particular visual referent to be described. To investigate structural guidance, we analyze the *constituent structure* of the sentences uttered. Specifically, we focus on noun and verb phrases, and calculate their constituent size as the number of content words (e.g., *the man* has size 1, *the tall man* has size 2). We obtain the constituent structure of a sentence using an automatic chunker (Daumé III & Marcu, 2005), which tags every word using a BIO encoding: the label B-X marks the beginning of phrase X, I-X marks an internal constituent of phrase X, and O marks words outside a phrase (see Appendix A for details). In order to simplify the interpretation of our results, we sum the frequency of beginning and internal constituent for each type of phrase, hence obtaining an aggregate measure rather than two distinct measures.

We also calculate the total number of constituents and control for its effects on all dependent measures reported in this study. We do this by residualizing it against the dependent measure under analysis in a simple linear regression model, e.g. $depM \sim NumConstituents$, using the `R` syntax, and we take the residuals obtained as the dependent measure for further inferential analysis.

Recall that each cue word ambiguously corresponded to two objects in the image. For the data analysis, we therefore needed to decide which of the two objects is referred to in the image description; this decision was made based on the transcripts of the speech produced by a participant for each trial. We mark as Primary the target objects that are spatially proximal to each other (e.g., in Figure 1, for the description *the man is writing on the clipboard*, the MAN and CLIPBOARD involved in this action are marked as primary). The remaining two target objects are marked as Secondary (in our example, this would be the MAN and CLIPBOARD not mentioned in the description). This distinction allows us to correctly assign visual responses to objects of interest (e.g., fixation latencies to the referent and its competitor.)[2] Two other cases can also arise, which we also mark: Both objects

[2]We include the Primary/Secondary distinction as a random variable in all mixed models concerning with latencies,

Category	Variable	Abbreviation
Oculomotor	Gaze Duration	Gaze
	Number of Fixated Objects	FixBefore/After
Conceptual	Animacy of Cue (Animate, Inanimate)	Cue
Perceptual	Scene Clutter (Minimal, Cluttered)	Clutter
Structural	Size of Constituent Type	NP, VP

Table 1
Independent variables analyzed in this experiment, with abbreviations used in the presentations of mixed effects models.

are mentioned in the sentence (e.g., *a man is writing on a clipboard while the other man is reading a letter* mentions both MEN), or the description is Ambiguous with respect to the object mentioned (e.g., *the man is sitting* could refer to either MAN). These two cases are, however, excluded only for the analysis of latencies, as a clear eye-voice relationship could not be established. In the Supplementary Material, we provide the reader with the complete list of descriptions produced in this study.

Our analyses also include oculo-motor variables to control for effects driven by mechanisms of visual attention that are independent of, or co-vary with, the effects brought on by our variables of interest. Oculo-motor variables will be discussed in more detail later in this section. In Table 1, we give an overview of all variables included as predictors in the mixed effects models reported in the result section.

Attentional Landscape and Turbulence

In this set of analyses, we quantify how fixations are distributed spatially, using the entropy of attentional landscapes (e.g., Pomplun, Ritter, & Velichkvosky, 1996; Henderson, 2003), and measure the complexity of the temporal sequence of fixated objects using the turbulence of scan patterns (Elzinga & Liefbroer, 2007).

An attentional landscape is a probability map of fixated locations computed by generating two-dimensional Gaussian distributions around the x/y coordinates of each fixation, with the height of the Gaussian weighted by fixation duration, and a radius (or standard deviation) of one degree of visual angle (roughly 20 pixels), to approximate the size of the fovea. A map is computed in this way for each subject on each trial and normalized to be a probability distribution, i.e., all values on the map sum to one. This then allows us to compute the global entropy of an attentional landscape, which intuitively represents the spread of information on the map: the higher the entropy, the more spread out are fixations on the scene. The entropy $H(L)$ of an attentional landscape is calculated as:

$$H(L) = \sum_{x,y} p(L_{x,y}) \log_2 p(L_{x,y}) \tag{1}$$

where $p(L_{x,y})$ is the normalized fixation probability at a point x, y in the landscape L. We examine how the entropy of fixations is mediated before and during production by the different types of

and all four cases for the analysis of fixation distribution and scan-pattern complexity.

guidance. On this measure, we mainly expect perceptual and conceptual guidance to play a role: attentional landscapes do not represent the sequential aspect of eye-movements, hence structural effects cannot be captured adequately.

The complexity of visual responses is measured on scan patterns using *turbulence*. Originally used in population research to quantify the complexity of life trajectories (Elzinga & Liefbroer, 2007), when applied to a scan pattern, turbulence is a measure of its variance, calculated by taking into account the number of unique objects fixated, the number of sub-sequences that could be generated from the scan pattern, and the relative duration of fixations. Intuitively, a scan pattern that contains only two objects (e.g., MAN-L/400 ms, CLIPBOARD-R/50 ms),[3] one of which is fixated most of the time, is less turbulent than a scan pattern in which a larger number of objects are fixated for an equal amount of time (MAN-L/200 ms, MAN-R/150 ms, CLIPBOARD-L/200 ms). Appendix B contains a detailed explanation of how turbulence is computed.

Latencies of Looks and Fixation Duration

The analysis of latencies differs across language production studies in the literature: latencies can be measured either at the onset or at the offset of the fixation on the object of interest (see Griffin and Davison (2011) for a review). In the present study, we measure the latency of the last look to the object of interest (EVS) at the offset of the fixation, as this is the moment when attention shifts away from it. By looking at the offset, we can confidently exclude cases where the fixation on the object crosses the point of mention, i.e., the duration of the last fixation overlaps with the time of mention. When looking at the trend of EVS latencies, we rely on a relative analysis, similar to that of Kuchinksy et al. (2011). In such an analysis, we consider all trials and account for the temporal variability of mention by normalizing latency and duration according to the time of mentioning. For example, if the referent was mentioned at 3500 ms, and the last look to its associated object was at 1500 ms for a duration of 300 ms, then we have a relative latency of 1500/3500 = 0.42 and a duration of 300/3500 = 0.08. All values of latency and duration then range between zero and one, with one indicating the longest latency. We also examined the onset of first look prior the mention of the referent, and the results are reported in the Supplementary Material.

In order to precisely estimate EVS latencies, our analyses include the following oculomotor variables: gaze duration of the fixation associated with the latency and number of objects fixated before latency is measured (from onset of trial until the last fixation on target object), as well as after (from the last fixation on target until mention). These are illustrated in Figure 2. We take into account these variables as they indicate how much visual information has been sampled before, during, and after the fixation latency for the object of interest is measured.

Inferential Statistics and Model Selection

Along with descriptive statistics, we report inferential statistics computed using linear mixed effects (LME) models (Pinheiro & Bates, 2000; Baayen, Davidson, & Bates, 2008). In each case, we model the dependent measure (e.g., latency for the EVS analysis) by using a set of perceptual, conceptual, and structural predictors (refer to Table 1 for an exhaustive list of predictors). As said above, all dependent measures are residualized for total number of constituents. The random variables considered are Participant, Item, and Mention, where Mention is a categorical variable indicating the type of mentioning scenario (Primary, Secondary, Both and Ambiguous), as described above. For

[3] We use -L and -R to denote the leftmost and rightmost object in case of ambiguous objects.

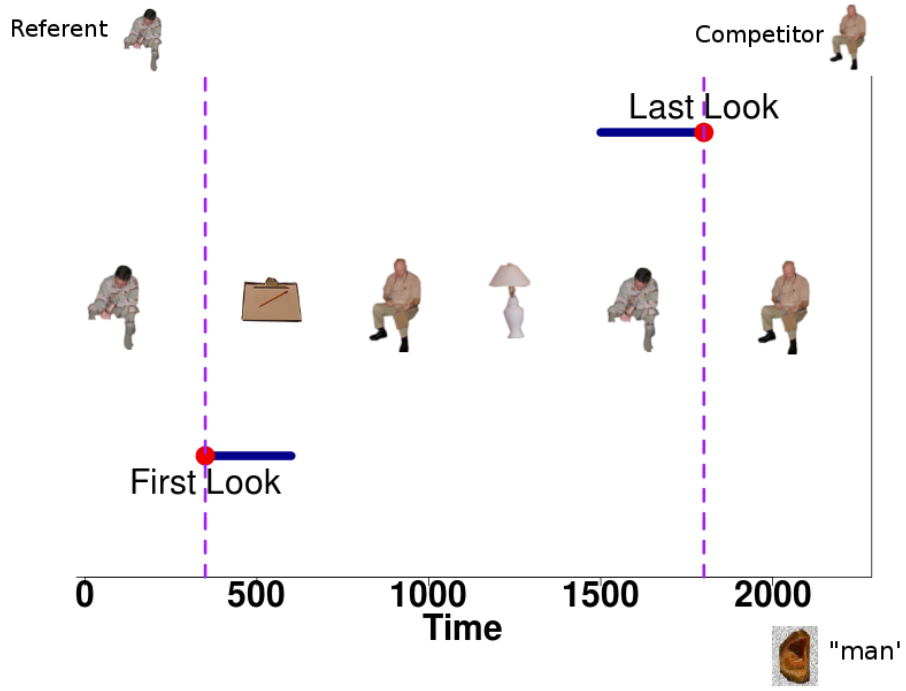


Figure 2. Schematic representation of the oculomotor variables involved in the latency of first look (measured at the onset of the fixation) and last look (measured at the offset of fixation) to the object of interest relative to the onset of mention (e.g., MAN). Dependent measure: latency, red dot; oculomotor predictors: number of objects fixated before and after the latency, duration of the fixation on the target (blue line). In the paper, we report only results for the last look, and refer the reader to the Supplementary Material for additional results on the first look.

the analysis of latencies, the variable will be reduced to the two levels (Primary, Secondary), as the data for the other two classes is not considered in the analysis. To reduce co-linearity, all factors were centered.

We perform model selection to obtain a minimal mixed effect model. It follows a best-path step-wise forward selection procedure, which compares nested models based on log-likelihood model fit. This model selection procedure is shown by Barr, Levy, Scheepers, and Tily (2013) to give a rate of Type-1 error comparable to a model with a maximal random structure. Models are fitted using maximum-likelihood estimation. We start with an empty model, to which we add the random factors, evaluated as intercepts. Once the random factors are included we continue with the fixed effects. For each fixed predictor, we evaluate whether random slopes have to be included. The inclusion of random slopes accounts for variability within different grouping of the random effects (e.g., participants). We add predictors one at time, ordered by the improvement they bring to the model fit. Only predictors that significantly improve model fit are retained. Significance is determined using a χ^2 test which compares the log-likelihood of the model before and after adding the new factor. Our model selection algorithm avoids the inclusion of fixed or random effects and associated slopes if this would be illegitimate, either because of contrast coding (e.g., an interaction of contrasted variables), or because of the structure of the experimental design. Our procedure

Predictor	β	*SE*	*t*	*p*
Intercept	0.007	0.729	0.01	0.9
Region	0.209	0.061	3.447	0.001
Clutter	0.124	0.018	6.933	0.0001
Cue	0	0.03	0.007	0.9
Region:Cue	0.107	0.037	2.916	0.004
Region:Clutter	0.08	0.036	2.252	0.02

Formula: (1 | Participant) + (1 | Item) + (1 | Mention) + Region + (0 + Region | Participant) + (0 + Region | Mention) + (0 + Region | Item) + Clutter + Cue + (0 + Cue | Item) + (0 + Cue | Participant) + Region:Cue + Region:Clutter

Table 2

Mixed model analysis of the Entropy of attentional landscapes. The predictors included in the model are: Clutter (minimal: −.5; cluttered: .5); Cue (inanimate: −.5; animate: .5); and Region (before speech onset: −.5; during speech: .5). The random variables included are: Participant (24), Item (24) and Mention (4)

returns the maximal random effect structure justified by the data and provides a principled way of selecting this structure, i.e., there is no need to commit, a priori, to a specific maximal random effect structure.

In the results tables, we report the coefficients and standard errors of the LME models, and derive *p*-values from the *t*-values (also reported) for each of the predictors in the model. Moreover, in the table, we also report the formula of model selected, using the R's `lme4` syntax.

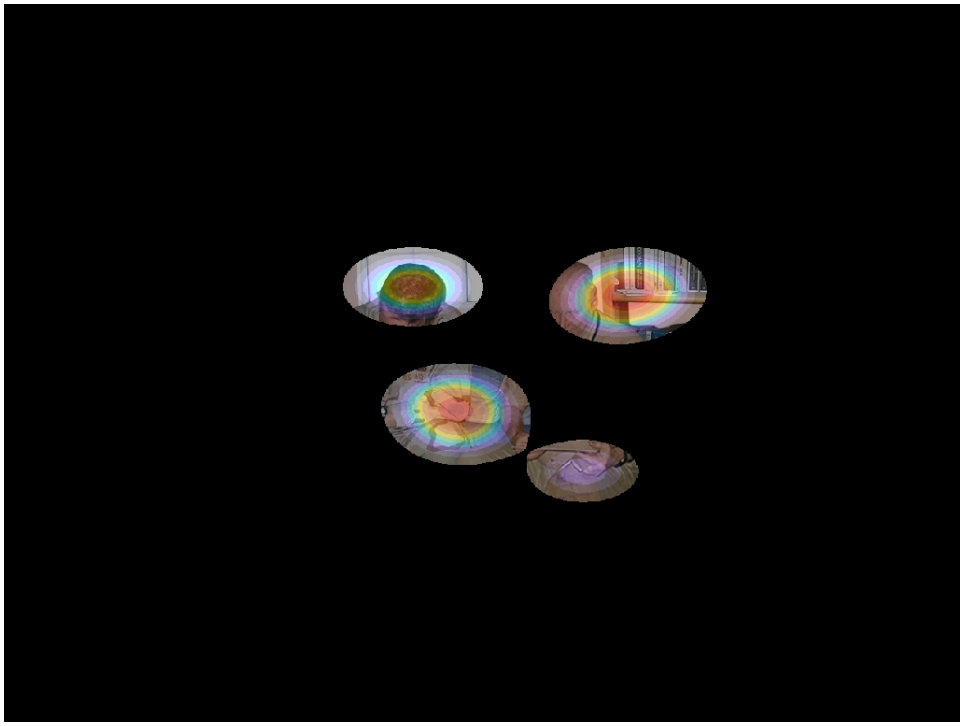
Results

What Drives Attentional Complexity

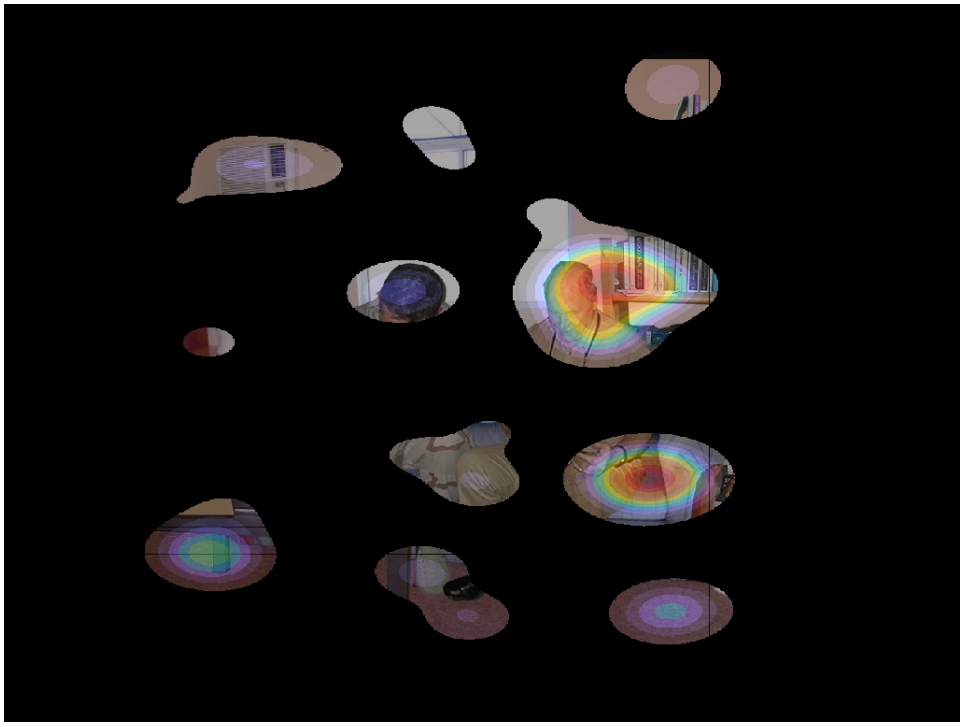
First, we look broadly at how the spatial distribution and the sequential structure of visual responses are modulated by the different components of guidance in the two phases of before and during sentence processing.

In this analysis, we are interested on the influence exerted by different forms of guidance on visual attention, regardless which referent is actually mentioned. Figure 3 illustrates an example of attentional landscape before and during the description of an animate referent, showing that attention is more spread out during speech. To quantify how factors such as perceptual (scene Clutter), conceptual (Cue animacy), and structural (number of NP and VP phrases) predictors, as well as the phase of the task (Region; before or during speech) modulate attentional spread, we applied a mixed model analysis to the entropy of attentional landscapes.

Table 2 shows the mean and standard error of the estimated coefficients for the predictors included during model selection, and their statistical significance. As expected, we observe a main effect of temporal Region: during speech, the entropy is higher than before speech onset. During surface realization, visual attention samples more widely the scene to better contextualize the ongoing description. Entropy also increases when the scene is cluttered. The more visual material is



(a) Before Speech Onset



(b) During Speech

Figure 3. Example of attentional landscapes while describing the animate cue (*man*) in a cluttered scene, before speech onset (upper panel) and during speech (lower panel). The heat map represents the probability of fixation, from low (blue) to high (red).

available, the more spread out on the scene attention becomes. Likewise, in order to generate a rich description, a large amount of visual information has to be sourced by the sentence processor for encoding. Of particular interest are the interactions. We find that entropy increases more during speech than before speech, especially when: (1) the scene is cluttered, and (2) the cue is animate.

These interactions highlight the important role of both perceptual and conceptual mechanisms when visual attention co-occur synchronously with sentence processing during a naturalistic production task. A cluttered scene contains more visual information that could be used for linguistic encoding than a minimal scene. This is especially crucial during speech, i.e., when visually attended information is realized as a sentence. Moreover, in a description, an animate object is usually contextualized within the scene as a whole, whereas an inanimate object tends to be described in terms of the spatial relation with other objects. Therefore, when describing an animate object, visual attention needs wider access to scene information, leading to increased entropy.

Entropy quantifies the spatial distribution of fixations, thus collapsing the sequential component of visual responses, which is crucial when looking for fine-grained effects associated with syntactic structure. Thus, in order to better understand effects of structural guidance, we next examine how the complexity of a scan pattern, operationalized as turbulence, varies with the syntactic structure of the associated sentence (expressed by constituent types NP and VP)

In Figure 4 we illustrate how turbulence varies in the two phases of the task (Before, During) as a function of the number of verb phrases (VP) and noun phrases (NP) produced. Recall that the turbulence of a scan pattern is an aggregate measure of its variability (see section Data Analysis above and Appendix B for details).

From the visualization, it is clear that scan-pattern complexity is greater during speech than before it, and that this effect is strongly modulated by the number of nouns and verbs produced (refer to Table 3 for coefficients and significance). So, in order to test whether NP and VP significantly differed in their slope on scan-pattern turbulence, we conducted an additional ANOVA analysis on two linear models predicting the turbulence of a scan-pattern given the type of phrase (NP, VP) and its frequency, focusing on the data from the during speaking phase. In particular, We compared a linear model including the slope Phrase:Frequency, with another one without it.

The results show a significance main effect of phrase, whereby VP generates higher turbulence than NP ($\beta = 3.03, t(4.75), p < 0.0001$), a main effect of frequency, whereby the higher the frequency, the more the turbulence, ($\beta = 1.08, t(15.78), p < 0.0001$), and a significant interaction phrase and frequency ($\beta = 0.74, t(3.70), p < 0.0002$), whereby the complexity on scan-pattern increases more for a unit increase of the frequency of verbs, than for unit increase of the frequency of nouns. ANOVA between the two models, in fact, confirms that the model with the interaction phrase and frequency is statistically better than the model without ($p < 0.0001$).

This analysis demonstrates that the strength of structural guidance is mediated by the type of phrasal constituent involved. Noun phrases visually correspond to single objects (e.g., *the man* and MAN), whereas verb phrases do not usually have a simple denotation but are associated with clusters of related objects (e.g., *eats an apple* is associated with APPLE, MOUTH, HAND). Hence, with increasing verbal material to be visually integrated, more objects have to be explored. This contrasts with nominal material, where there is often a simpler one-to-one mapping.

It is also interesting to observe that the complexity is not uniquely modulated by structural factors; rather, we observe several significant interactions occurring between the three forms of guidance. In particular, cluttered scenes are associated with more complex scan-patterns, especially when the cued target is animate (two-way interaction Clutter:Cue). This result indicates that the

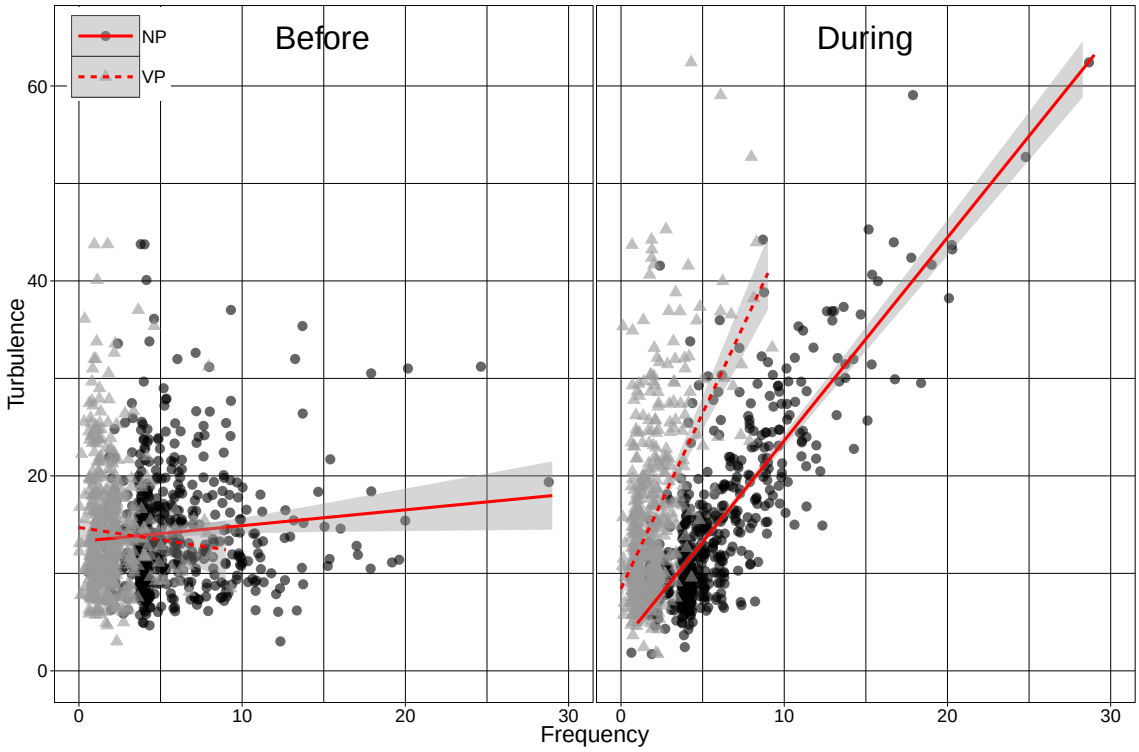


Figure 4. Scatter-plot of Turbulence (y-axis) as a function of the Frequency of Noun and Verb Phrases (x-axis), respectively marked using point and line type (NP: circle-solid, VP: triangle-dashed), divided in the two phases of the task (Before, left panel; During, right panel). The lines represents the mean estimates (and standard errors as shaded bands) of a generalized linear model fit to the data.

description of an animate target is contextualized more broadly with the rest of the scene, than an inanimate target which, instead, is described relative to specific landmarks.

Cluttered scenes also correlate with richer verbal structure during speaking (three-ways interaction Region:Clutter:VP). This result highlights that structural and perceptual interactively mediate attentional guidance during language production. In fact, sentences with rich verbal information can only be generated when there is sufficient visual information available, which is reflected by more complex eye-movement responses.

We also find a main effect of Cue, whereby animate targets have a lower turbulence than inanimate targets, especially before speaking, as indicated by the two-ways interaction Region:Cue. During speaking instead, an animate target triggers higher turbulence than an inanimate target, especially for sentences richer in nouns (three-ways interaction Region:Cue:NP). In Table 4, we provide some examples of descriptions produced by the participants together with their entropy and scan-pattern turbulence score.

Taken together, these results suggest that structural guidance follows mechanisms of referential integration, e.g., the eye-voice span. However, the granularity of this guidance depends on the syntactic categories of the sentence being constructed, with each syntactic category interfacing

Predictor	β	*SE*	*t*	*p*
Intercept	-0.093	0.492	-0.189	0.8
Region	1.03	0.757	1.361	0.1
Clutter	0.741	0.302	2.455	0.01
Cue	-0.631	0.308	-2.05	0.04
NP	0.067	0.067	1.009	0.3
VP	-0.106	0.15	-0.708	0.4
Region:NP	1.497	0.129	11.623	0.0001
Region:VP	1.305	0.297	4.390	0.0001
Cue:NP	-0.262	0.088	-2.964	0.003
Region:Cue	1.433	0.615	2.328	0.02
Clutter:Cue	1.328	0.608	2.183	0.02
Region:Cue:NP	0.664	0.177	3.761	0.0001
Region:Clutter:VP	1.164	0.471	2.471	0.01

Formula: (1 | Participant) + Region + (0 + Region | Participant) + Clutter + Cue + NP + VP + Region:NP + Region:VP + Cue:NP + Region:Cue + Clutter:Cue + Region:Cue:NP + Region:Clutter:VP

Table 3
Mixed model analysis of the Turbulence of the scan-patterns. The predictors included in the model are: Clutter (minimal: −.5; cluttered: .5); Cue (inanimate: −.5; animate: .5); Region (before speech onset: −.5; during speech: .5); NP and VP (frequency count variables). The random variables included are: Participant (24)

with particular aspects of real-world information. Moreover, all forms of guidance co-exist, and mutually interact, a result that challenges a strict interpretation of structural guidance during situated language production.

Competition and Referential Ambiguity

The next set of results concerns visual responses related to the mention of the cued target. We concentrate on the issue of referential ambiguity (recall that the cue can refer to two distinct visual objects in the scene) and show how the different forms of guidance modulate ambiguity resolution. The measures reported for this result section are calculated only on those cases when the Primary and Secondary referent are mentioned (55% of the cases; 322/576), independently for Animate and Inanimate targets, i.e., we exclude the cases Ambiguous (18%; 103/576) and Both (26%; 151/576), for which it is not possible to establish an unambiguous competitor. In a large proportion (26%) of the cases, the referent is not looked at, possibly indicating an effect of para-foveal or peripheral preview, i.e., the target is not foveated (see Appendix C for more details).

In Table 5, we report the percentage of times the referent and the competitor are fixated in

Condition	Complexity	Sentence			Turbulence		Entropy	
			NP	VP	Bef	Aft	Bef	Aft
Minimal-Animate	High	The girl holding the teddy bear is sitting on the bed and the girl in the nightie is next to the bed	13	4	15.44	36.88	11.08	11.85
	Low	The girl hugged her teddy	4	1	9.03	4.99	10.74	11.09
Cluttered-Animate	High	There is a girl sitting on her bed hugging her teddy bear and another little girl next to her she looks like she wants the teddy bear	15	7	10.75	36.55	11.01	11.84
	Low	The man is sitting and the man is standing	4	4	4.97	12.10	10.28	11.12
Minimal-Inanimate	High	There is some sort of reception counter and a telephone on the reception counter next to a man there is another telephone on a pedestal to his left	18	2	30.52	42.37	11.53	11.55
	Low	The shoe is on the floor	4	1	23.69	4.98	11.48	10.01
Clutter-Inanimate	High	There is one shoe in a filing box next to a woman on a bed in a bedroom and another shoe on the floor near another woman also in the bedroom	20	1	31.01	43.68	11.68	12.27
	Low	The towel is folded on the bed	4	2	17.31	6.35	11.42	10.70

Table 4

Example of sentences produced by the participants across experimental conditions for two classes of visual response complexity (High, Low). Together with the sentence, the table provides the frequency of noun and verb constituent (NP, VP), as well as the complexity values for Turbulence, Entropy in the two phases of the task (Before and During speech onset).

		Animate		**Inanimate**	
	Overall	Cluttered	Minimal	Cluttered	Minimal
Referent	74% (239/322)	28% (69/239)	33% (80/239)	17% (41/239)	20% (49/239)
Competitor	61% (199/322)	32% (64/199)	35% (70/199)	14% (28/199)	18% (37/199)

Table 5

Percentage of times the referent object and the competitor object are fixated (Overall) when the description is unambiguous (Primary and Secondary). Fixations are broken down by experimental condition: Clutter (Cluttered, Minimal) and Cue (Animate, Inanimate).

	Animate		Inanimate	
	Cluttered	Minimal	Cluttered	Minimal
Before	0.06	0.03	0.46	0.76
During	0.93	0.80	0.07	0.29

Table 6
Relative difference in percentage fixation to the referent object and its competitor for the two phases of the production task (Before speech onset and During speech). The ratios are calculated relative to the competitor. Thus, positive values indicate a higher probability of looking at the referent than its competitor and vice-versa. Values close to 0 indicate equal probability of looking at either referent. The ratios are broken down by experimental condition: Clutter (Cluttered, Minimal) and Cue (Animate, Inanimate).

unambiguous descriptions, i.e. when only one of the two referents was mentioned. We break down these figures by visual clutter and animacy of the cue. We find that referents are fixated overall more often than competitors, and the ratio (74/61) reveals a dominance of about 20%. The fact that the percentage of looks is above 50% for both referents indicates the prominent guiding effect of the cue and the strong resulting competition when the alternative referent is often evaluated against the referent selected for naming.

In particular, when the cue is animate, both target and competitor are fixated more often than when the cue is inanimate, especially when the scene has a minimal visual clutter. These results are consistent with work on visual search, which has found that animate objects are identified more easily than inanimate objects (Fletcher-Watson et al., 2008), and that more clutter makes target identification more difficult (Henderson et al., 2009). This effect is especially strong for inanimate targets, which are harder to distinguish from scene clutter than animate ones. The result of competition between referentially ambiguous objects begs for closer examination of the time-course during which competition develops. In particular, we ask whether the competition between the mentioned referent and its competitor is stronger before speech begins, or while it unfolds.

To quantify the amount of competition present, Table 6 reports the difference between the percentage of looks to the referent and its competitor. We present these ratios for the two phases of the production task (before speech onset and during speech), across the four experimental conditions. We find that before production, the competition between referent and competitor is very high, i.e., the difference index is close to 0, especially when the cue is animate. However, once overt speech begins, we find the referent is looked at twice as often as the competitor. This effect is modulated by both animacy of the cue, and visual clutter. Animate cues trigger a higher competition than inanimate cues, especially before speech. For inanimate cues, we observe a stronger focus on the referent already before speech, especially when the scene has minimal clutter. However, during speech, referent and competitor appear to be fixated with similar probabilities, especially for cluttered scenes. These results are in line with existing studies in visual search, which have found inanimate targets to be less likely to be fixated in cluttered scenes. However, the indexes reported above, describing changes in probability of looking at either referent in ambiguous scenes, bring also new insights on the processes underlying ambiguity resolution: first, they elucidate that referential ambiguity resolution takes place mostly before speech, and secondly, they show that the

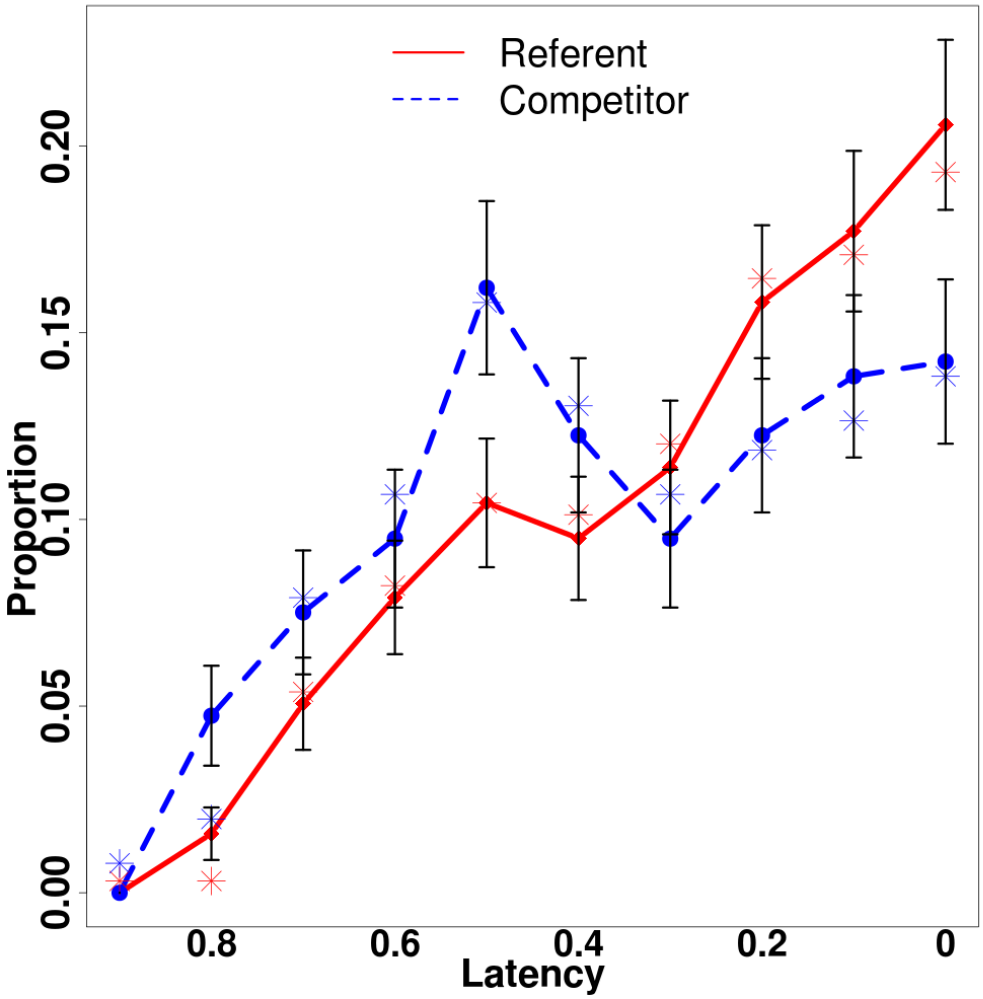


Figure 5. Proportion of EVS latencies at different temporal blocks relative to the onset of the mentioned target. The latency measures the time elapsed from the end of the last fixation to the object (referent or competitor) until it is mentioned. Red is the referent, blue the competitor. The asterisks indicate the LME model fit.

conceptual information carried by the cue and the global scene information strongly modulate the amount of competition between multiple referents.

Besides probabilities of looks, another important aspect of referential competition regards specific latencies of looks to each object, i.e. their timing. In particular, we look at the (EVS) latency, i.e., the latency of the last look before mention, which is calculated only on those cases where the mentioned object can be identified unambiguously; and report results on the latency of first look in the Supplementary Material.

We find that the referent has a relative mean latency of 0.32, which is shorter than that of its competitor (0.37). Gaze duration on referents (0.09) is very similar to the one on competitors (0.10). To enable us to understand the temporal distribution of EVS latencies, Figure 5 shows the proportion

of latencies at different temporal intervals (0.1 each) relative to the time of mention. Proportions increase towards the mention for both the referent and the competitor, i.e., the closer to the mention we are, the more gazes are associated with the referent object, a pattern also observed by Qu and Chai (2008). In line with Fukumura and Van Gompel (2011), our results clearly confirm referential competition, with the competitor displaying both an early and a late peak. We can conjecture that during the early peak (about 0.5 relative latency), visual attention evaluates whether the competitor is a better linguistic candidate, whereas the late peak (around 0.1 relative latency) indicates a final visual check on the competitor before the referent is encoded linguistically. This result corroborates with the fact that a stronger competition is observed before speech begins (refer to Table 6 for details).

The EVS proportions therefore provide evidence that visual attention is not just guided by structural properties of the encoding; rather, attention makes also active use of perceptual information in the scene context. More specifically, the presence of a competitor activates comparative processes, despite the fact that only one object is eventually selected for mention. A strict interpretation of structural guidance would predict looks only on the mentioned object.

Table 7 displays the coefficients for a mixed model that predicts relative EVS latency based on perceptual, structural, conceptual, and oculo-motor factors. We find EVS to be significantly modulated by the number of objects fixated before and after it. In particular, the more objects are fixated after the EVS, i.e., the last look, the longer the latency, whereas the more objects are fixated before it, the shorter the latency. These effects have an intuitive mechanistic explanation: if many objects are fixated before the last look to the object of interest, then it is more likely that this look occurs closer to mention. Likewise, if many objects are fixated after it, then it is more likely that this look occurred early on. The significant effect of Gaze duration follows the same logic: the longer the gaze, the shorter the latency. A longer gaze implies that more visual information about the referent has been retrieved. This adds an implicit cost to the latency, i.e., it pushes it closer to mention.

Besides being influenced by oculomotor predictors, the latency of EVS is modulated by structural predictors. In particular, we find that an increased number of noun phrases (NP) results into shorter EVS latency. However, this effect is modulated by the the number of objects fixated before the EVS. In fact, we find that the more the objects fixated, the more the nouns to be produced, the longer the latency of the EVS (two-way interaction FixBefore:NP). When more nominal material is planned for encoding, more objects are looked at beyond the referent object to situate it in the scene. Thus, a description containing a rich nominal structure, which correlates with a high number of objects being fixated prior to EVS, demands the referent to be inspected as early as possible. This mechanism makes the referent less likely to be looked at its linguistic mention.

Finally, we also observe an interaction between scene clutter and animacy of the cue: an inanimate target in a cluttered scene has a longer EVS than an animate target. An inanimate referent is usually described in the context of spatial landmarks surrounding it. A cluttered scene increases the number of possible landmarks that are evaluated for encoding, hence anticipating the last look to the referent.

This second set of results on latencies corroborate our results about spatial and temporal aspects of attentional complexity reported above. In particular, we observed perceptual and conceptual guidance to mediate attention distribution over the scene. The reason is that perceptual and conceptual guidance exert a direct impact on the visual search routines used to identify the target object and retrieve information about the surrounding objects to be mentioned in the description. Attentional spread was predicted by perceptual and conceptual information. This claim is further supported by

Predictor	β	*SE*	*t*	*p*
Intercept	0.004	0.009	0.513	0.6
FixAfter	0.613	0.040	15.365	0.0001
FixBefore	-0.366	0.041	-8.992	0.0001
Gaze	-0.412	0.076	-5.447	0.0001
Clutter	0.008	0.010	0.767	0.4
NP	-0.007	0.003	-2.453	0.01
Cue	0.006	0.016	0.394	0.6
Clutter:Cue	-0.060	0.020	-3.006	0.003
FixAfter:Gaze	-0.975	0.388	-2.514	0.01
FixBefore:NP	0.025	0.012	2.140	0.03

Formula: (1 | Item) + FixAfter + FixBefore + Gaze + Clutter + NP + Cue + (0 + Cue | Item) + Clutter:Cue + FixAfter:Gaze + FixBefore:NP

Table 7

Mixed model analysis of the eye-voice-span (EVS). The predictors included in the model are: Clutter (minimal: −.5; cluttered: .5); Cue (inanimate: −.5; animate: .5); number of Object fixated, prior and after the last look (FixBefore, FixAfter); Gaze duration of fixation; and NP (frequency count). The random variables included is: Item (24)

the effect of perceptual guidance on the latency of first look, which is when search routines of target identification are still operating (the reader is referred to the Supplementary Material for an analysis for the latency to the first look to the mentioned referent and its ambiguous competitor). However, once the message has to be verbally realized, visual attention becomes more and more dependent on the sequential processing of linguistic information. Here, structural guidance plays a major role in the visual responses at this time. Indeed, our analysis of scan pattern complexity corroborates the EVS latency results, as both analyses show that structural properties of the sentence, such as its syntactic composition, have a direct effect on the complexity of the associated visual responses both at the local level of the latency, as well as on the entire scan pattern. Most importantly, all three forms of guidance examined in this study co-exist and mutually interact to drive language production in naturalistic scenes.

Discussion

The main goal of this eye-tracking study was to demonstrate that different forms of guidance are actively involved in the control of visual attention during language production. We investigated three forms of guidance: perceptual, conceptual, and structural. These three forms of guidance have previously been studied mostly in isolation. Furthermore, compared to prior work using the visual world paradigm, we increased the ecological validity of both the linguistic task, by using an image description task constrained only by the cued target, and of the visual aspects of the task, by using

photographic scenes. This resulted in a more realistic study of the dynamics of the cross-modal processing of visual and linguistic information. Specifically, we manipulated perceptual guidance through the visual clutter in the scene, conceptual guidance by cuing either animate or inanimate objects, and structural guidance by examining the syntactic structure of the sentences produced, which we represented as the number of constituents, and the frequency of phrase labels of noun and verb phrases.

We performed two different sets of analyses. We initially considered the broad patterns of attentional complexity and scan pattern complexity and investigated the influence exerted by the different forms of guidance on the global allocation of visual attention, both before speech onset and during speech onset. Then, we focused on the mention of the cued target to investigate the issue of referential ambiguity.

Consistent with the visual cognition literature, we found that the entropy of the attentional landscape is greater when the scene is visually cluttered, especially during speech. Access to scene information was also found to depend on the animacy of the object to be described: animate objects generate a wider sampling during speech compared to inanimate objects. Animate objects are conceptually more accessible than inanimate ones (Levelt et al., 1999), which has been shown to facilitate message planning before speech (Coco & Keller, 2009). However, such accessibility also implies more interaction of the animate referent with its context: during speech, more visual information has to be evaluated to contextualize the description of the animate object within the scene.

To capture in more detail the effect of structural guidance, we then looked at visual responses in their sequential form, i.e., as scan patterns. Representing the structural complexity of a sentence as the the frequency of VP and NP constituents (see Table 1) and using turbulence as the complexity measure for visual responses we found different patterns before speech onset and during speech. Intuitively, a scan pattern with many unique objects, all fixed for the same amount of time, is more turbulent than a scan pattern with fewer objects, some which are fixated longer than others. Our hypothesis that scan pattern complexity should correlate with sentence complexity, and that the exact type of syntactic constituents should play a role, was verified.

We found that more complex syntactic structures are associated with more turbulent scan patterns. This effect is stronger during sentence production, which is the phase of the task demanding more synchronization between visual and linguistic information. We also found verbal material to have a larger effect on turbulence during sentence production than nominal material. The explanation is that verbs are visually associated with clusters of related objects. Hence, if more verbal material has to be encoded, then more visual objects require inspection.

Moreover, we found interesting interactions of structural guidance with perceptual and conceptual guidance. A cluttered scene was associated with high frequency of verb phrases, which in turn, resulted into more complex scan-pattern responses. Animate targets were instead associated with high frequency of noun phrases, and again, this was reflected by more complex scan-pattern responses. This result highlights an inter-dependent relationship between forms of visual guidance during situated language production.

In the second set of analyses, we looked at the latencies of the last look to the referent and its ambiguous competitor relative to the onset of mention, (e.g., Griffin & Bock, 2000; Qu & Chai, 2008). Specifically, we assessed whether the referent object and its ambiguous competitor were both attended before mention, and to uncover how guidance mediates such looks. A strict interpretation of the structural guidance hypothesis predicts that only the mentioned referent is targeted by visual attention, i.e., we would expect EVS for the referent and its competitor to be clearly different. How-

ever, it is known that the presence of a competitor in the visual context impacts on the production of target expressions (Fukumura et al., 2010). This makes the opposite prediction: the referent and its competitor should have similar EVS profiles.

Our results provide support for the latter prediction: we find that both the referent object and its competitor display similar time courses, i.e., the alignment between gaze and name increases closer to mention (in line with Qu & Chai, 2008). This finding rules out a strict interpretation of structural guidance as proposed by Kuchinksy et al. (2011). In our study, language production seems to be guided by structural information in the form of a specific linguistic referent; however, the visual presence of a competitor modulates the production process. Based on the EVS time course, we were able to identify distinct peaks in the latency of the referent and the competitor, indicating the presence of two phases: (1) ambiguity resolution at the early stage of EVS, and (2) gaze-to-name binding, occurring closer to mention.

Crucially, EVS is also mediated by an interaction between perceptual and conceptual factors. In particular, inanimate referents display longer latency of EVS in cluttered scenes. We argued that a cluttered scene contains more landmarks that could be used to ground the description of an inanimate referent; which implies an earlier look to it prior to its linguistic mention.

Unlike previous work, our analysis of latencies took into account the impact of oculomotor control variables. In particular, we looked at how the amount of visual information sampled before or after fixating the referent. This was done to ensure a more accurate estimation of the EVS. Indeed, we demonstrated the importance of such variables by showing that the more objects are inspected prior to the last fixation, the shorter the latency; the more objects are fixated after it, the longer the latency. We additionally found an effect of gaze duration on EVS: the longer the gaze, the shorter the EVS latency.

Our results show also an effect of structural guidance on EVS latency. In particular, we find that an increase in the amount of nominal material in the utterance leads to a shorter latency of mention. The demand of planning many NPs forces visual attention to time-lock more strongly the gaze-to-name mapping. However, we also found that this effect is modulated by oculo-motor variables. In particular, description rich in noun phrases, which correlates with a high number of objects being fixated prior to EVS, triggered longer EVS to the mentioned referent. We argued that more nouns to be produced imply more objects to be fixated, forcing attention to evaluate the referent being produced as early as possible, and move on onto inspecting other object, prior to its mention. These result pin down more precisely how the syntactic component of structural guidance modulates EVS latency, and how it interacts with other key guidance components. Moreover, this way of approaching structural guidance generalizes to other situated language production tasks, as it does not rely on formulaic language use (e.g., telling the time).

To summarize, our results not only confirm the existence of perceptual guidance, previously disputed in the literature, they also bring evidence to support a new form of guidance, viz., conceptual guidance. Furthermore, our results clearly demonstrate the interaction between all three forms of guidance, thus paving the way for detailed theoretical accounts that predict how the language production system integrates perceptual, structural, and conceptual information before and during speaking. Finally, our work significantly extends previous research by providing an analytical framework that makes it possible to quantify the structural complexity involved in both sentences and scan patterns. We also uncovered a novel set of phenomena that surface when sentence production is examined in a more naturalistic context.

General Discussion

Everyday language is often situated in a visual context, such as when we give directions on a map, explain the function of a device, or tell the time. In such tasks, linguistic and visual processing have to be synchronized. Describing images, for example, requires visual attention to recover information about objects, properties, and events from a visual scene, while language processing has to integrate and verbalize this information and produce a sentence.

The interaction between visual attention and sentence processing could potentially take place in a number of ways, and a wide variety of factors have been identified, both linguistic (e.g., Meyer et al., 1998; Griffin & Bock, 2000; Bock et al., 2003; Kuchinksy et al., 2011) and perceptual (e.g., Gleitman et al., 2007; Papafragou et al., 2008; Myachykov et al., 2011). Language production is sensitive to simple perceptual events (e.g., a light flash at the target location, Gleitman et al., 2007), but also interacts with complex perceptual information, such as the type of motion depicted (influencing whether manner or path verbs are produced (Papafragou et al., 2008), or the complexity of the depicted event (influencing grammatical role assignment and linear position, Myachykov et al., 2011). These findings have been used to argue for an interactive account of visual guidance, according to which linguistic and perceptual processing work in tandem to guide visual attention during language production (Gleitman et al., 2007). On the other hand, there is considerable evidence for the existence of an eye-voice span: objects are fixated approximately 900 ms before being mentioned (e.g., Griffin & Bock, 2000; Bock et al., 2003; Qu & Chai, 2008). This has been used to argue that visual attention is mostly guided by structural constraints on sentence processing, a view that is supported by Kuchinksy et al.'s (2011) study, which showed that structural, but not perceptual information, guides attention in formulaic tasks such as telling the time.

The present article reported an experiment which brings together structural and perceptual factors, with the aim of unraveling how these factors guide visual attention. In a cued language production task using photo-realistic scenes, we demonstrated that there are at least three forms of visual guidance (perceptual, conceptual, and structural) which interact with each other. Which type of guidance is active strongly depends on which stage the language production system is in: when focusing on the cued object, we find that perceptual guidance dominates until this object is fixated, while structural guidance dominates after it has been fixated, but before it is mentioned (i.e., during the eye-voice span). If we consider the time course of production as a whole, then we find that perceptual, structural, and conceptual guidance are all active, but the three forms of guidance exert a stronger effect during speaking, compared to before speech onset.

The aim of the present study was not only to clarify the role of perceptual and structural guidance, but also to demonstrate that this is not a simple dichotomy: visual attention in situated production is guided by a range of additional factors. Specifically, we introduced the notion of conceptual guidance, i.e., guidance provided by non-linguistic properties of the referents being processed. The category of a referent to be visually searched (e.g., Henderson & Hollingworth, 1999; Zelinsky & Schmidt, 2009) or to be linguistically described (e.g., McDonald et al., 1993; Branigan et al., 2008) is already known to mediate visual and linguistic responses. In our study, we manipulated conceptual guidance by cuing descriptions of either animate or inanimate visual referents.

We investigated the interaction of the three types of guidance in four different analyses (entropy of attentional landscapes, turbulence of scan patterns, latency of first fixation and eye-voice-span). Contrary to claims of Kuchinksy et al. (2011), we demonstrated that structural guidance alone is not sufficient to explain the mechanism underlying the eye-voice span, as we found that referen-

tially ambiguous objects display a similar EVS, despite the fact that only one of them is mentioned. Moreover, we found that EVS was mediated by oculomotor mechanisms, which can only be examined in naturalistic visual contexts, such as photo-realistic scenes.

The use of photo-realistic scenes is therefore a methodological innovation of our study. It makes visual responses more likely to be guided by naturalistic mechanisms of scene understanding, rather than uniquely depending on linguistic processing. The use of naturalistic scenes contrasts with current practice in psycholinguistic research on situated language processing, which has almost exclusively relied on object arrays or clip-art scenes (but see Andersson, Ferreira, & Henderson, 2011).

In order to fully explore our data, we went beyond standard analyses (e.g., latencies), and applied analytical tools from visual cognition (attentional landscapes, Pomplun et al., 1996) and population research (turbulence, Elzinga & Liefbroer, 2007) to quantify the impact of different types of guidance on visual attention. In particular, when investigating the turbulence of scan patterns, we found that perceptual guidance (clutter), structural guidance (number of constituents), and conceptual guidance (animacy) are all active, again contradicting a structural-guidance-only view.

Our results rely on general definitions for both perceptual and structural guidance, and we offer ways of quantifying their impact on visual attention. In particular, in line with research in visual cognition (e.g., Itti & Koch, 2000), we define perceptual guidance in terms of low-level visual information and measure it as clutter (Rosenholtz, Li, & Nakano, 2007). We operationalize structural guidance in terms of syntactic information and measure it on the basis of the constituent structure of an utterance. This differs from Kuchinksy et al. (2011), who define perceptual guidance as experience with reading clocks, while structural guidance is operationalized in terms of the instructions given to participants. Both definitions are highly task-dependent and thus unlikely to generalize, as task is known to play an important role in the allocation of visual attention (e.g., Rothkopf, Ballard, & Hayhoe, 2007; Castelhano, Mack, & Henderson, 2009). Our definition of structural guidance also makes it possible to link syntax and eye-movements directly. We demonstrated for the first time that each syntactic type carries a precise cost in terms of the complexity of visual responses. In particular, noun phrases, which allow a one-to-one mapping with visual objects (e.g., *the apple* relates to the object APPLE), are less costly than verb phrases, which instead relate to a cluster of visual objects (e.g., *eats an apple* relates to the objects APPLE, MOUTH, HAND). This result paves the way for future research in which the relation between linguistic and visual complexity can be further elucidated, perhaps by representing sentences as syntactic trees rather than as sequences of part-of-speech tags.

Overall, our results support an interactive account of situated language production, where different cross-modal mechanisms (perceptual, conceptual, and structural) are synchronized to produce visual and linguistic representations. Such synchronization occurs by means of cross-modal alignment of the information provided by sentences and scan patterns. This result has practical applications, as we show in Coco and Keller (2012): The close alignment of linguistic structure and visual attention can be exploited to predict which sentence a participant will utter based on the scan pattern they follow. Such predictions draw upon the fact that similar sentences tend to occur with similar scan patterns, where both sentence and scan pattern similarity can be quantified using sequential similarity measures. On the theoretical level, the results of Coco and Keller (2012) raise intriguing questions regarding the linguistic (e.g., semantic) and visual (e.g., saliency), mechanisms involved in cross-modal coordination during situated language processing, while also providing a framework in which such mechanisms can be investigated (viz., using cross-modal similarity).

It is also important to draw a connection to research on situated language comprehension. This literature has emphasized that visually presented referential information impacts sentence processing at different levels, including phonological (e.g., beaker vs. beetle, Allopenna, Magnuson, & Tanenhaus, 1998), semantic (e.g., piano vs. trumpet, Huettig & Altmann, 2005), and visual (e.g., rope vs. snake, Dahan & Tanenhaus, 2005). Our study suggests that language production is influenced by a similar set of factors (structural, conceptual, and perceptual). However, the direction of multi-modal integration in production is reversed, i.e., scene understanding has to happen before constraints supplied by visual attention are applied to sentence production (Altmann & Kamide, 1999; Knoeferle & Crocker, 2007; Kukona, Fang, Aicher, Chen, & Magnuson, 2011). Future research is needed to investigate the similarities and differences between situated comprehension and production in more detail, perhaps in a dialogue setting in which both processes happen at the same time.

References

Allopenna, P., Magnuson, J., & Tanenhaus, M. (1998). Tracking the time course of spoken word recognition: evidence for continuous mapping models. *Journal of Memory and Language*, *38*, 419–439.

Altmann, G., & Kamide, Y. (1999). Incremental interpretation at verbs: restricting the domain of subsequent reference. *Cognition*, *73*, 247–264.

Andersson, R., Ferreira, F., & Henderson, J. (2011). I see what you are saying: The integration of complex speech and scenes during language comprehension. *Acta Psychologica*, *137*, 208–216.

Arnold, J., & Griffin, Z. (2007). The effect of additional characters on choice of referring expression: everything counts. *Journal of Memory and Language*, *56*, 521–536.

Baayen, R., Davidson, D., & Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390-412.

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.

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

Branigan, H., Pickering, M., & Tanaka, M. (2008). Contribution of animacy to grammatical function assignment and word order during production. *Lingua*, *2*(118), 172-189.

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.

Castelhano, M., Mack, M., & Henderson, J. (2009). Viewing task influences eye-movement control during active scene perception. *Journal of Vision*(9), 1–15.

Coco, M. I., & Keller, F. (2009). The impact of visual information on referent assignment in sentence production. In *In N.A. Taatgen and H. van Rijn (Eds.), Proceedings of the 31th Annual Conference of the Cognitive Science Society, Amsterdam.*

Coco, M. I., & Keller, F. (2012). Scan patterns predict sentence production in the cross-modal processing of visual scenes. *Cognitive Science*, *36*(7), 1204–1223.

Dahan, D., & Tanenhaus, M. (2005). Looking at the rope when looking for the snake: Conceptually mediated eye movements during spoken-word recognition. *Psychological Bulletin and Review*, *12*, 455–459.

Daumé III, H., & Marcu, D. (2005). Learning as search optimization: Approximate large margin methods for structured prediction. In *International Conference on Machine Learning (ICML).*

Elazary, L., & Itti, L. (2008). Interesting objects are visually salient. *Journal of Vision*, *8*(14:18), 1–15.

Elzinga, C., & Liefbroer, A. (2007). Destandardization of the life course: A cross-national comparison using sequence analysis. *European Journal of Population*, *23*(3-4), 225–250.

Ferreira, V., Slevc, L., & Rogers, E. (2007). How do speakers avoid ambiguous linguistic expressions? *Cognition*, *96*, 263–284.

Findlay, J., & Gilchrist, I. (2001). Visual attention: The active vision perspective. In *M. Jenkins & L. Harris (Eds.), Vision and Attention* (pp. 83–103). Springer-Verlag, New York.

Fletcher-Watson, S., Findlay, J., Leekam, S., & Benson, V. (2008). Rapid detection of person information in a naturalistic scene. *Perception*, *37*(4), 571–583.

Frank, M., Vul, E., & Johnson, S. (2009). Development of infants' attention to faces during the first year. *Cognition*(110), 160–170.

Fukumura, K., & Van Gompel, R. (2011). The effects of animacy in the choice of referring expressions. *Language and Cognitive Processes*, *26*, 1472–1504.

Fukumura, K., Van Gompel, R., & Pickering, M. J. (2010). The use of visual context during the production of referring expressions. *Quarterly Journal of Experimental Psychology*, *63*, 1700–1715.

Gabadinho, A., Ritschard, G., Müller, N., & Studer, M. (2011). Analyzing and visualizing state sequences in R with TraMineR. *Journal of Statistical Software*, *40*, 1–37.

Gleitman, L., January, D., Nappa, R., & Trueswell, J. (2007). On the give and take between event apprehension and utterance formulation. *Journal of Memory and Language*, *57*, 544–569.

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

Griffin, Z., & Davison, J. (2011). A technical introduction to using speakers' eye movements to study language. In *The mental lexicon* (Vol. 6, pp. 55–82). John Benjamins Publishing Company.

Henderson, J. (2003). Human gaze control during real-world scene perception. *Trends in Cognitive Sciences*, *7*, 498-504.

Henderson, J., Chanceaux, M., & Smith, T. J. (2009). The influence of clutter on real-world scene search: Evidence from search efficiency and eye movements. *Journal of Vision*, *9(1)*(32), 1–8.

Henderson, J., & Hollingworth, A. (1999). High-level scene perception. *Annual Review of Psychology*, *50*, 243-271.

Huettig, F., & Altmann, G. (2005). Word meaning and the control of eye fixation: Semantic competitor effects and the visual world paradigm. *Cognition*, *96*(1), B23-B32.

Hwang, A., Wang, H., & Pomplun, M. (2011). Semantic guidance of eye movements in real-world scenes. *Vision Research*, *51*, 1192–1205.

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, *40*(10-12), 1489–1506.

Kennedy, A., & Pynte, J. (2005). Parafoveal-on-foveal effects in normal reading. *Vision Research*, *45*, 153–168.

Knoeferle, P., & Crocker, M. (2007). The influence of recent scene events on spoken comprehension: Evidence from eye movements. *Journal of Memory and Language*, *57*(4), 519–543.

Kuchinksy, S., Bock, K., & Irwin, D. (2011). Reversing the hands of time: Changing the mapping from seeing to saying. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*, 748–756.

Kukona, A., Fang, S., Aicher, K., Chen, H., & Magnuson, J. (2011). The time course of anticipatory constraint integration. *Cognition*, *119*, 23–42.

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

Levin, H., & Buckler-Addis, A. (1979). *Eye-voice span*. The MIT Press Classics Series.

McDonald, J., Bock, J., & Kelly, M. (1993). Word and world order: semantic, phonological and metrical determinants of serial position. *Cognitive Psychology*, *25*(2), 188Ã¢ÂĂÂŞ230.

Meyer, A., Sleiderink, A., & Levelt, W. (1998). Viewing and naming objects: eye movements during noun phrase production. *Cognition*, *9(11)*(8), 1–13.

Myachykov, A., Thompson, D., Scheepers, C., & Garrod, S. (2011). Visual attention and structural choice in sentence production across languages. *Language And Linguistic Compass*, *5*(2), 95–107.

Nelson, W., & Loftus, G. (1980). The functional visual field during picture viewing. *Journal of Experimental Psychology: Human Learning and Memory*, *7*, 369–376.

Noton, D., & Stark, L. (1971). Eye movements and visual perception. *Scientific American*, *224*(1), 34–43.

Nuthmann, A., & Henderson, J. (2010). Object-based attentional selection in scene viewing. *Journal of Vision*, *10*(8:20), 1–20.

Papafragou, A., Hulbert, J., & Trueswell, J. (2008). Does language guide event perception? *Cognition*, *108*, 155–184.

Parkhursta, D., Lawb, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, *42*(1), 107–123.

Pinheiro, J., & Bates, D. (2000). *Mixed-effects models in s and s-plus*. Statistics and Computing Series, Springer-Verlag, New York, NY.

Pomplun, M., Ritter, H., & Velichkvosky, B. (1996). Disambiguating complex visual information: Towards communication of personal views of a scene. *Perception*, *25*, 931–948.

Prat-Sala, M., & Branigan, H. (2000). Discourse constraints on syntactic processing in language production: a cross-linguistic study in english and spanish. *Journal of Memory and Language*, *42*, 168–182.

Pynte, J., New, B., & Kennedy, A. (2008). On-line contextual influences during reading normal text: A multiple-regression analysis. *Vision Research*, *48*(21), 2172–2183.

Qu, S., & Chai, J. (2008). Incorporating temporal and semantic information with eye gaze for automatic word acquisition in multimodal conversational systems. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP). Honolulu.*

Qu, S., & Chai, J. (2010). User language behavior, domain knowledge, and conversation context in automatic word acquisition for situated dialogue. *Journal of Artificial Intelligence Research*, *37*, 247-277.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*(3), 372–422.

Rosenholtz, R., Li, Y., & Nakano, L. (2007). Measuring visual clutter. *Journal of Vision*, *7*, 1-22.

Rosenholtz, R., Mansfield, J., & Jin, Z. (2005). Feature congestion, a measure of display clutter. *SIGCHI*, 761-770.

Rothkopf, C. A., Ballard, D. H., & Hayhoe, M. M. (2007). Task and context determine where you look. *Journal of Vision*, *7*(14), 1–20.

Russell, B., Torralba, A., Murphy, K., & Freeman, W. (2008). Labelme: a database and web-based tool for image annotation. *International Journal of Computer Vision*, *77*(1-3), 151-173.

Tanenhaus, M., Spivey-Knowlton, M., Eberhard, K., & Sedivy, J. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*(268), 632–634.

Underwood, G., & Foulsham, T. (2006). Visual saliency and semantic incongruency influence eye movements when inspecting pictures. *The Quarterly Journal of Experimental Psychology*, *59*, 2031–2038.

Zelinsky, G., & Schmidt, J. (2009). An effect of referential scene constraint on search implies scene segmentation. *Visual Cognition*, *17*(6), 1004-1028.

Appendix A
Syntactic Chunking

In order to quantify syntactic guidance, the sentences produced in our experiment were decomposed into syntactic chunks. Syntactic chunks are an intermediate representation between simple parts of speech and a full parse, which gives us basic structural information about the complexity of a sentence.

Chunking was performed automatically using the TagChunk system developed by Daumé III and Marcu (2005), which performs combined part of speech tagging and syntactic chunking. It assigns syntactic labels to a sequence of words using the BIO encoding, in which the beginning of phrase X (e.g., noun phrase or NP) is tagged B-X (e.g., B-NP), the non-beginning (inside) of the X phrase is tagged I-X (e.g., I-NP), and any word that is not in a phrase is tagged O (outside). The TagChunk systems achieves a chunking accuracy of 97.4% on the CoNLL 2000 data set (8,936 training sentences, and 2,012 test sentences). As an example, consider a description of the scene in Figure 1, and its chunked version:

(1) There is a clipboard sitting on a coffee table and another clipboard next to it on which a US army man is writing.

(2) [B-NP There] [B-VP is] [B-NP a [I-NP clipboard]] [B-VP sitting] [B-PP on] [B-NP a [I-NP coffee [I-NP table]]] [B-O and] [B-NP another [I-NP clipboard]] [B-ADJP next] [B-PP to] [B-NP it] [B-PP on] [B-NP which] [B-NP a [I-NP US [I-NP army [I-NP man]]]] [B-VP is [I-VP writing]].

Based on the TagChunk output, we can calculate the frequency of each syntactic type and the total number of constituents (e.g., the frequency of B-NP is 7 in the above sentence), which we use as factors in the mixed models reported in the main text.

Appendix B
Turbulence

The concept of turbulence of a sequence comes from research in population research, where the goal is to quantify the complexity of different life trajectories (e.g., Single, Married, Married with Children, sequence: S–M–MC, vs. Single only, sequence: S; Elzinga & Liefbroer, 2007). Turbulence is a composite measure calculated by integrating information about: (1) number of states a sequence is composed of, (2) their relative duration, and (3) the number of unique subsequences that can be derived from it.

Suppose we have three different scan patterns (x = MAN-R, MAN-L, CLIPBOARD, WINDOW, y = MAN-R, CLIPBOARD, MAN-R, MAN-L, z = MAN-R, CLIPBOARD, MAN-L, MAN-R) each of which consists of four fixated objects. Intuitively, x is more turbulent than y and z, as in x all fixated objects are different and inspected only once. When we compare y and z instead, we find that they have the same number of uniquely fixated objects, but in z we find that more objects are fixated before looking back at MAN-R. Therefore, z can be considered to be more turbulent than y, as more events have occurred before MAN-R is inspected again.

A combinatorial implication of this reasoning is that more distinct subsequences, denoted as $\phi(x)$, can be generated from z than from y. In fact, when computing the number of distinct subsequences for the three scan patterns, we find that $\phi(x) = 16 > \phi(z) = 15 > \phi(y) = 14$.[4] Each state of a sequence, however, is often associated with a certain duration. Let's assume MAN-R is fixated for

[4]The logarithm of $\phi(x)$ is taken to avoid large numbers as $\phi(x)$ increases exponentially with increasing x.

200 ms before looking at another object. If we include fixation duration on x, we will obtain something like MAN-R/200, MAN-L/200, CLIPBOARD/200, WINDOW/200, and its turbulence increases when the variance duration between states decreases. So, x has a relative high turbulence (variance of 0 across states), compared to a scan pattern such as MAN-R/400, MAN-L/50, CLIPBOARD/75, MAN-L/50, where the duration is concentrated on a single state.

These intuitions can be formalized by defining the turbulence of a sequence as (Elzinga & Liefbroer, 2007):

$$T(x) = \log_2 \left(\phi(x) \frac{s^2_{t,max}(x) + 1}{s^2_t(x) + 1} \right) \tag{2}$$

where $\phi(x)$ denotes the number of distinct subsequences, s^2_t is the variance of the state-durations and $s^2_{t,max}$ the maximum of that variance given the total duration of the sequence, calculated as $s^2_{t,max} = (n-1)(1-\bar{t})^2$ with n denoting the number of states and $\bar{t}$ the average state duration of the sequence x. To compute the turbulence for the analyses in the main text, we utilized the `R`-package `TraMine`, a toolbox developed by Gabadinho, Ritschard, Müller, and Studer (2011) to perform categorical sequence analysis.

Appendix C
Distance from Mentioned, but not Fixated, Objects

Since the distance from a fixation is known to have processing implications (Nelson & Loftus, 1980), we performed an additional analysis in which we looked at the closest fixation to the referent when is not mentioned, and calculate the distance of this fixation from the object centroid. We find that on average, fixations are at 11.65 ± 5.04 degree of visual angle from the object centroid. This indicates that peripheral vision is sufficient to identify and select the object for mentioning: for parafoveal effects, the object should be fixated within 4–5 degrees of visual angle. Moreover, a mixed model analysis reveals that the distance varies with the amount of visual clutter and the animacy of the target. In particular, an animate object tends to have a smaller visual distance ($\beta_{Animate} = -3.12$; $p < 0.05$), especially in a scene with minimal clutter ($\beta_{Animate:Minimal} = -6.25$; $p < 0.01$). This indicates that foveating the object is not an obligatory element of mentioning.