Quarterly Journal of Experimental Psychology 67:6, 1096-1120, 2014.

The Interplay of Bottom-Up and Top-Down Mechanisms in Visual Guidance During Object Naming

Moreno I. Coco, George L. Malcolm, 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

moreno.cocoi@gmail.com

Abstract

An ongoing issue in visual cognition concerns the roles played by low- and high-level information in guiding visual attention, with current research remaining inconclusive about the interaction between the two. In this study, we bring fresh evidence into this long-standing debate by investigating visual saliency and contextual congruency during object naming (Experiment 1), a task in which visual processing interacts with language processing. We then compare the results of this experiments to data of a memorization task using the same stimuli (Experiment 2). In Experiment 1, we find that both saliency and congruency influence visual and naming responses, and interact with linguistic factors. In particular, incongruent objects are fixated later and less often than congruent

ones. However, saliency is a significant predictor of object naming, with salient objects being named earlier in a trial. Furthermore, the saliency and congruency of a named object interact with the lexical frequency of the associated word and mediate the time-course of fixations at naming. In Experiment 2, we find a similar overall pattern in the eye-movement responses, but only the congruency of the target is a significant predictor, with incongruent targets fixated less often than congruent targets. Crucially, this finding contrasts with claims in the literature that incongruent objects are more informative than congruent objects by deviating from scene context, and hence need a longer processing. Overall, this study suggests that different sources of information are interactively used to guide visual attention on the targets to be named, and raises new questions for existing theories of visual attention.

Keywords: Eye-movements; object naming; scene understanding; cross-modal processing; visual guidance.

Introduction

Research in visual cognition has demonstrated that the allocation of visual attention is influenced by both low-level properties of the scene percept and high-level conceptual knowledge related to it. Low-level properties are perceptual features of a scene (e.g., color) and can be quantified using measures such as saliency (Itti & Koch, 2000). Conversely, high-level knowledge refers to semantic properties of the scene, such as the category that objects belong to (Zelinsky & Schmidt, 2009) and, more generally, contextual information conveyed by a scene (Brooks, Rasmussen, & Hollingworth, 2010; Eckstein, Drescher, & Shimozaki, 2006). Depending on the type of task performed, visual attention might rely more on stimulus- or knowledge-based features (Hayhoe & Ballard, 2005; Einhauser, Rutishauser, & Koch, 2008). In particular, saliency is a good predictor of visual attention when there are no specific target objects set by the task. During free-viewing (Parkhurst, Law, & Niebur, 2002) or memorization (Underwood & Foulsham, 2006), saliency can predict up to the first five fixations better than chance. However, during tasks where visual attention is cued to a specific target (e.g., visual search), saliency is a poor predictor of performance (Henderson, Brockmole, Castelhano, & Mack, 2007; Einhauser, Rutishauser & Koch, 2008; Henderson, Malcolm, & Schandl, 2009). In this case, low-level information has to be modulated by contextual information to make better predictions of fixation location (e.g., PEDESTRIANS are often found on

the STREET, Torralba, Oliva, Castelhano, & Henderson, 2006, Ehinger, Hidalgo-Sotelo, Torralba & Oliva, 2009).

Understanding the bi-directional interplay between saliency and context has critical implications for developing and refining theories of visual cognition. Two theoretical positions have been advocated in the literature so far, representing extreme ends of the bottom-up/top-down debate of visual attention guidance: First, the stimulus-driven approach based on a bottom-up architecture of the visual system assumes that target selection is driven by the low-level features of the incoming scene percept (Walther & Koch, 2006). In this framework, the recognition of objects becomes collateral to the fact that salient regions correlate with the presence of objects (Elazary & Itti, 2008).

Alternatively, the knowledge-based approach, exemplified by the Cognitive Relevance Framework (Henderson et al., 2009), assumes that locations are targeted according to their contextual relevance in relation to the task performed. Cognitive relevance is derived from target information in conjunction with the scene context in which the target occurs (Malcolm & Henderson, 2010; Castelhano & Heaven, 2010).

Research continues to elucidate the interplay between saliency and contextual information in gaze guidance (Foulsham & Underwood, 2011), and the definition and applicability of both concepts are subject to revision (Baluch & Itti, 2011); especially

when these factors are observed in the context of real-world behavior (Tatler, Hayhoe, Land, & Ballard, 2011).

Real-world contexts are often dynamic, and it is clear that a combination of bottom-up and top-down information is used to optimize the allocation of visual attention by minimizing uncertainty, hence maximizing the likelihood of achieving the goals set by the task.

However, the debate surrounding saliency and context stems from conflicting evidence found in studies that investigate their interaction and their impact on visual attention during static-picture viewing.

For example, Underwood et al. (2008) investigated the bottom-up/top-down interplay by utilizing an odd-object task in which participants were given a scene containing a contextually incongruous object (i.e., an object that violates the contextual expectation of a scene, such as a COW on a SKI SLOPE). In recognition tasks, odd objects are identified less readily than congruous objects (e.g., Davenport & Potter, 2004), but, paradoxically, various studies indicate that incongruent objects attract fixations earlier (e.g, Loftus & Mackworth, 1978; Underwood, Humphrey, & Cross, 2007; Bonitz & Gordon, 2008; but see De Graef, Christiaens, and Ydewalle, 1990; Henderson, Weeks & Hollingworth, 1999; and Vo & Henderson, 2009, 2011).

Thus, despite the inhibited recognition performance of incongruent objects, their contextual incompatibility seems to attract visual attention earlier than congruent objects. In order to explain this paradox, Underwood et al. (2008) argued that in the studies which found effects of contextual incongruence, incongruent objects might have been more salient than congruous ones, meaning that early fixations to the target object would be due to low- rather than high-level features. This intuition stemmed from previous work by the authors (Underwood & Foulsham 2006), where saliency and congruency were also manipulated but not in a controlled and systematic way[1]. In order to formally test this hypothesis, Underwood, et al. (2008) asked participants to complete a comparative visual search task in which they had to spot a changed object in two otherwise identical, side-by-side scenes. Critically, the saliency and contextual congruency of the target object was manipulated independently, and the time to first fixate the object was measured. The results showed contextually incongruous objects (e.g., a tin of tomatoes in a washing machine) were fixated earlier, but saliency had no effect.

Interestingly, these results are at odds with what the same authors observed in previous work. Underwood and Foulsham (2006) found that semantic incongruency boosted the first fixation to the incongruent object only during memorization, but not in search. Again, only in memorization, they also found an effect of visual saliency: a salient object

1 Only congruency was tested through a rating task, where participants were asked to rate how likely (1-9) an object was to occur with the rest of the scene.

was fixated earlier than a non-salient object. The effect of saliency, however, did not interact with semantic incongruency in any of the two tasks. Underwood et al. (2008) did not discuss the difference with their previous work, and instead speculated that the paradox between the visual system's inhibition for recognizing incongruous objects, but apparent facilitation in having visual attention drawn to these same objects, could be due to participants' pre-attentively[2] locating objects in a scene whose low-level features deviate from the overall gist. For example, Li, Van Rullen, Koch, and Perona (2002) found that certain objects could be identified without attention; however, later research by Evans and Treisman (2005) argued that the recognition without attention found by Li et al. (2002) might be due to the target objects' intermediate distinctive features which are detected without having to operate attentional binding. Moreover, within the context of feature-driven attentional allocation, the idea that there is a pre-attentive stage of attention appears to be contentious. Studies on predictive processes of visual search, in fact, suggest that initial processing might be driven by re-configuration strategies tuning the visual system to prioritize features of the incoming stimuli that are useful to perform the task at hand (e.g., Di Lollo, Kawahara, Zuvic, & Visser, 2001; Enns and Lleras, 2008). Furthermore, problems with defining a pre-attentive early stage of visual processing emerge also when looking at electro-physiology data, where areas other than V1 are found to be active already after 30 ms from the onset of the visual percept (see Foxe and Simpson, 2002).

2 The authors used the term pre-attentively when referring to the early stage of gist processing. In this paper, we will not make a distinction between different attentional stages.

An alternative way to explain the mechanisms of contextual expectation guiding visual attention, which does not distinguish between attentive stages of processing, is predictive coding (see Clark, 2012, for a synthesis of this perspective). The predictive coding account explains the interaction between top-down representational knowledge and bottom-up perceptual information in terms of error correction (e.g., Rao & Ballard, 1999, Hinton, 2007, Friston 2010). At the core of predictive coding is a hierarchical generative network model, in which the expected information, e.g., the fact that a CUP is usually found on a TABLE, is actively utilized to interpret and integrate incoming information. If the incoming information matches, the representation is consolidated; if it does not match, the representation is updated using the prediction error. Within this framework, an incongruent object violating contextual expectations, e.g., a CUP in a BATHTUB, triggers a prediction error. This error is used to update our contextual expectations, i.e., CUP can be also found in BATHTUBS. For visual attention, this would imply longer search latencies for the indentification of incongruent objects, as there are many contextually congruent objects interfering with its identification, i.e., attention will be allocated on the most likely objects available in the context. Such interference between contextually congruent objects would also mediate how much it would be attended, e.g., total fixation duration, observed on the incongruent object. In particular, a shorter total fixation duration is expected on the incongruent object. As attentional resources are primarily allocated to contextually congruent objects, an incongruent object would be less likely re-fixated.

Our study focuses on the interaction between saliency and congruency, rather than on temporal stages of visual attention, so we will not enter directly into the controversy about pre-attentive processes when interpreting our results. However, we will draw parallels between what we observe and the predicting coding framework, as it offers a natural way to interpret our results.

In the present study, we aim to recreate the manipulation of saliency and contextual congruency used in Underwood et al.'s (2008) study in order to address the bottom-up/top-down issue without some of the limitations of Underwood et al.'s results: The absence of saliency effects in their study might be due to the comparative nature of the search task, which encouraged participants to judge the contextual similarity of the two scenes rather than rely on differences in low-level information in order to identify the changed object. Furthermore, this effect could have been reinforced by the co-presence of both scenes, which led participants to develop a precise comparative scanning strategy (Gajewski & Henderson, 2005; Underwood, 2009).

Here we utilize an object naming task, in which participants had to name five objects within a scene image (Experiment 1), in order to investigate how attentional guidance is modulated by contextual and image features when search objects are naming targets. Moreover, we compare the eye-movement responses in this task with responses observed

during memorization (Experiment 2), which is a task commonly used in the literature on contextual effects in visual attention (e.g., Henderson, et. al., 1999, Underwood & Foulsham, 2006).

In the object naming task, mechanisms of linguistic encoding exploit the available visual information to select naming targets. This makes it possible to observe how low- and high-level information of the scene is attended and used to generate the naming sequences. The naming task has also other methodological advantages over previously used tasks. First, it does not require two scenes to be displayed side-by-side, thus we avoid inducing a comparative scanning strategy. Second, object naming overcomes another important limitation of Underwood et al. (2008) and other related studies in the literature: saliency and context have always been studied in purely visual tasks, in which only mechanisms of visual attention are actively engaged. In many realistic cognitive tasks, however, visual attention has to cooperate with other modalities (motor control, auditory processing, language processing) to achieve specific goals (e.g., Hayhoe & Land, 2005). This cooperation inevitably must draw upon both low- and high-level visual information; an interaction of the two types of information, and other information involved (linguistic in our study) is therefore more likely to manifest.

In language production tasks, visual attention retrieves information to be used when uttering words or sentences. The relevant linguistic material, such as the nouns referring

to the visual objects, is selected based upon the scene information provided by the visual system. Note that we are not making the claim that visual attention is directly involved in the retrieval of lexical information, but rather pointing out that visual attention is used to operate the first selection of the objects to be named; and that this selection is modulated by linguistic properties of the objects being attended. We therefore assume a cross-modal architecture in which information is accessed, shared and exchanged synchronously across different modalities. Moreover, when we discuss visual attention and language processing, we are not ascribing to them any goal or intention, rather we elucidate their cross-modal interaction.

Given this architecture and the information flow it entails, it seems reasonable to assume that low- and high-level visual mechanisms interact closely in a language production task such as object naming. This motivates our use of object naming in the present paper as a means of uncovering the interaction between saliency and contextual congruency; and to examine the role played by language processing in it.

Naming is a well-studied task in psycholinguistics, where it is used to investigate the linguistic mechanisms underlying language production (e.g., Levelt, Vorberg, Meyer, Pechmann, & Havinga, 1991; Meyer, Sleiderink, & Levelt, 1998; Griffin & Oppenheimer, 2006), as well as more generally to study the role of contextual information (e.g., Bartram, 1974; Snodgrass, 1980; Potter, Kroll, Yachzel, Carpenter, &

Sherman, 1986; Griffin & Bock, 1998; Damian, Vigliocco, & Levelt, 2001; Hocking, McMahon, & Zubicaray, 2009). Psycholinguistic results have given rise to an interactive account of naming, in which several types of constraints, linguistic and non-linguistic, mediate the selection of lexical items and influence the associated response times. On one hand, linguistic information such as lexical frequency (Meyer et al., 1998, Almeida, et al., 2007) or word length (Zelinsky & Murphy, 2000) modulates the associated gaze duration (less frequent or longer words correlate with longer gaze durations). On the other hand, the linguistic act of naming is constrained by the sentential context in which it is situated (Griffin & Bock, 1998), as well as by the semantics of surrounding objects (Damian et al., 2001; Hocking et al., 2009). Evidence for this is also found in more complex production tasks, such as scene description, which is influenced by low-level visual information (e.g., the cueing of a location through a brief flash, Gleitman, January, Nappa, & Trueswell, 2007) or by high-level semantic properties of objects (e.g., animacy, Coco & Keller, 2009).

However, none of the existing studies in either the visual cognition or the psycholinguistic literature directly investigates the interaction of visual saliency and contextual congruency during object naming. We use this simple linguistic task to shed light on the interaction between these factors, which are fundamental to scene understanding, and to investigate how linguistic mechanisms can mediate their access and use. The goal is to provide new evidence on how visual attention and language

processing exploit multi-modal information to build a common workspace, share resources, and draw joint inferences to tackle a contextually situated linguistic task. The insights obtained could form a first step towards an integrated theory of cross-modal processing that explains how multi-modal information is concurrently accessed and synchronously employed to perform cognitive tasks.

The current study investigates the interplay between low-level visual information (saliency) and high-level control (contextual congruency) during object naming (Experiment 1). In a follow-up experiment, we then compare the eye-movement responses observed during naming with the results of a memorization task (Experiment 2).

Experiment 1

Participants were asked to name five objects in a naturalistic scene. Each scene contained an object of interest whose saliency and congruency were manipulated. We hypothesize that both sources of information have to be integrated in order to optimally select objects to be named. In contrast to goal-directed tasks (e.g., search), a naming task is not cued to a single target object; instead, every object is a potential target. Therefore, the relevance of an object is determined solely by the viewer. The viewer can rely on an object's saliency, contextual congruency, or on both.

The viewer can also use linguistic knowledge in combination with these other two sources of information to select the object to name. Thus, a visually salient object could be fixated more often than a less salient one, just by virtue of its low-level properties (e.g., Elazary & Itti, 2008), leading to greater chance of it being named. This expectation would challenge a strict interpretation of goal-oriented theories of visual guidance (e.g., Einhauser, Rutishauser & Koch, 2008), in which objects are selected on the basis of the contextual information provided by the scene when a task demands top-down control.

However, naming objects demands an extensive processing of contextual information, as objects that are contextually similar are likely to be co-activated (Huettig and Altmann, 2005), hence facilitating their naming. Moreover, the importance of contextual co-occurrence is also observed as a facilitating factor during visual search (e.g., Mack and Eckstein, 2011, Hwang, Wang, & Pomplun, 2011).

For these reasons, we expect cognitive relevance, and the co-occurrence statistics on which it draws, to be a crucial component of naming. In particular, contextually congruent objects are more likely to be attended and consequently selected for naming.

Additionally, Underwood et al.'s (2008) account hypothesizes that the incongruent target object will have shorter latencies to first fixation than the congruent target objects, as it is

rapidly identified as deviant already during scene gist processing. In contrast, the interpretation of the Cognitive Relevance Framework predicts no difference in search latencies, i.e., the time until first fixation, between incongruent and congruent objects, as the scene gist will direct gaze without knowledge of each object's properties. A naming task benefits from contextual expectations, as words for semantically related objects are more easily retrieved from the lexicon and spelled out, than words for objects deviating from the context. Thus, we expect an incongruent object to be looked at, for the first time (search latency), later than a congruent object. Visual attention would be directed first to objects fitting the contextual congruency of the scene. This expectation would be compatible with the predictive coding framework: attention is initially captured by contextually congruent objects which interfere with one another in representational space, hence delaying the identification of the incongruent object. Visual saliency, however, is expected to compensate for this incongruency effect, by boosting search latency. So, a visually salient, incongruent object, should be identified at the same speed of a non-salient congruent object. As visual saliency boosts target identification, then a visually salient, and contextually congruent object should be the quickest to be looked at.

Moreover, if congruent objects are more likely targets, they should be attended for longer (e.g., total gaze) than incongruent objects. Contextually congruent objects compete for attentional resources, hence reducing the overall number of fixations on the incongruent object. Furthermore, the fact that the semantics of incongruent objects deviate from the

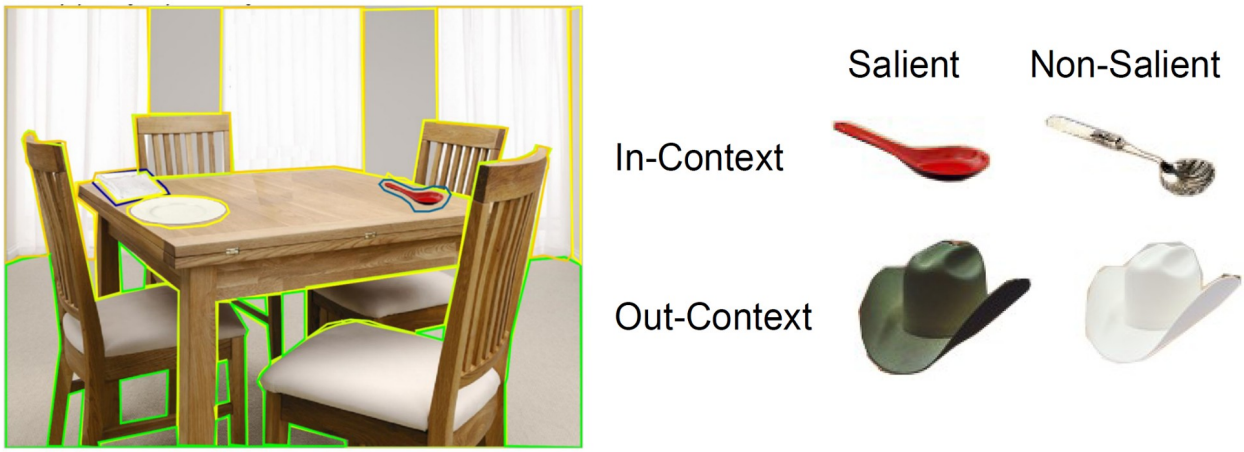
other objects in the context makes them more memorable, hence demanding less attentional processing. This effect, thus, is expected to carry over in the memorization task presented in Experiment 2. On this measure, we do not expect visual saliency to mediate its processing, as the low-level properties of the object do not modulate its semantic relation within the scene. Thus, a salient incongruent object would be looked at for the same amount of time, as a non-salient incongruent object.

Finally, since language processing is also directly implicated in the evaluation and linguistic selection of the target objects, we expect properties of this information stream to exert an influence on the responses observed, in line with previous work (e.g., Zelinsky and Murphy, 2000). Moreover, if the resolution of the task is really performed by drawing, cross-modally, on various sources of information, we expect saliency and congruency to interact with linguistic properties of the objects to be named, such as their lexical frequency, known to play a key role during naming (e.g., Levelt, Schriefers, Vorberg, Meyer, Pechmann, & Havinga, 1991).

Method

Participants were presented with photo-realistic scenes and asked to name five objects in each scene. They said the names of these objects out loud, and the speech recording was time-locked with participants' eye-movements. The scenes used in the experiment were

Figure 1. The left panel gives an example of an annotated photo-realistic scene used in the experiment. In this example, the object of interest is the LADLE or HAT. The right panel shows the experimental manipulation of Saliency and Congruency. The competitor objects are NAPKIN and PLATE.



created based on photographs obtained from the LabelMe database (Russell, Torralba, Murphy, & Freeman, 2008), which were selected to contain only inanimate objects[3]; refer to Figure 1 for an example of materials and experimental conditions.

In each scene, an object of interest and two competitors were inserted using Photoshop. Saliency (Salient, Non-Salient) and Congruency (In-Context, Out-of-Context) of the object of interest were manipulated. Each scene was fully annotated with polygons marking the outlines of objects. On average, scenes contained M = 14.07, SD = 6.17

3 Some scenes came from Google Images.

annotated objects. The polygons were used to map fixation coordinates into the corresponding objects.

The saliency of the object of interest was manipulated by changing its color, brightness/contrast, and hue/saturation with Photoshop. Moreover, the position of the object was sometimes slightly modified to boost its saliency with respect to the background. As it is known that visual saliency can be estimated in different ways, and that this can consequently result into different accuracy score when predicting visual responses across tasks (Borji, et al. 2012), we verified the effectiveness of the saliency manipulation using both Itti and Koch's (2000) and Torralba et al.'s (2006) models (in the latter case, only the saliency part of the model was used, not the context part). Note, moreover, that our manipulation strictly relates to a single object of interest and it is purely methodological: we do not make any claim about predicting visual responses in a scene based on saliency models. The object of interest was regarded as salient when the saliency values returned by both models were higher than those of the competitor objects. A t-test was used to confirm that salient objects ($M = 0.47$, $SD = 0.34$) had significantly higher saliency scores than non-salient objects ($M = 0.07$, $SD = 0.11$; $t(199) = 14.14$, $p < 0.0001$). In the non-salient condition, the saliency of the object of interest ($M = 0.03$, $SD = 0.1$) was significantly less than those of the competitors ($M = 0.07$, $SD = 0.09$; $t(211) = 2.5$, $p < 0.05$).

Contextual congruency was manipulated by replacing the object of interest with one that intuitively did not fit the scene context. The effectiveness of this manipulation was checked using Allison, Keller, and Coco's (2012) model of object context, which predicts how well an object fits with a set of other objects based on label co-occurrence counts derived from LabelMe. The model employs a distribution over the set of labels in the scene to generate a continuous measure of object fit in a scene. It explains the observation of sets of objects through latent scene types, which can be thought of as simple clusters of objects which are likely to co-occur.

The object of interest was regarded as incongruent when its context score according to the Allison et al. (2012) model was lower than that of the competitors. A t-test confirmed that incongruent objects had significantly lower context scores ($M = 0.16$, $SD = 0.1$) than congruent ones ($M = 0.64$, $SD = 0.2$; $t(43) = 10.95$, $p < 0.0001$).

The position of the object of interest was counterbalanced by rotating it in three different locations of the scene (Left, Middle, Right) for each condition, in order to account for possible directional and central biases (Tatler, 2007; Tatler & Vincent, 2008). Moreover, the counterbalancing significantly reduces the possibility that the effects observed are scene-specific. This resulted in a total of 12 versions of each scene (four conditions in three positions). A total of 28 different scenes were used for the experiment, as well as 28 fillers. The fillers were scenes drawn from the same database and also manipulated using

Photoshop to prevent participants from being able to distinguish fillers and experimental materials. In particular, we pasted a range of different objects, which varied both in saliency and in their congruency with the other objects in the scene. This was done to insure that the participants will not strategically detect the experimental trials by, for example, remembering that they always contained three pasted objects. Figure 1 gives an example of a scene used, together with the objects of interest corresponding to the four experimental conditions.

Twenty-four native speakers of English, all students of the University of Edinburgh, were each paid five pounds for taking part in the experiment. Informed consent was obtained from each participant prior to the experiment, and the task was explained using written instructions. The experiment took approximately 20 minutes to complete.

Each participant saw all fillers and each of the 28 experimental scenes in one condition. Items were distributed across participants in a Latin-square design to ensure that each condition was presented equally often to each participant. The order of fillers and experimental items was randomized. The items were preceded by four practice trials that served to familiarize participants with the task.

An Eyelink II head-mounted eye-tracker was used to monitor participants' eye-movements with a sampling rate of 500 Hz. Participants sat at approximately 50 cm from

a 21” Multiscan monitor, where scenes were presented at a resolution of 800 x 600 pixels, which subtend 31.26 degrees of visual angle on the horizontal axis, and 25.19 degrees on the vertical one. The object of interest, on which saliency and congruency were manipulated, subtended an average of 3.97 degrees of visual angle on the horizontal axis, and 3.20 degrees on the vertical one. Participants’ speech was recorded with a lapel microphone.

Only the dominant eye was tracked. At the beginning of each trial, participants looked at a fixation cross on the center of the screen, and drift correction was performed. The scene then appeared at which point participants were free to move their eyes; 1500 ms later an audible beep was played, indicating that they could start naming objects. Participants were told simply to name objects as quickly or slowly as they wanted, as long as they waited until after the beep. This was to dissuade participants from naming the first five objects they saw and instead ponder more carefully the five objects they felt were to be named. There was no time limit for the trial duration and to pass to the next trial participants pressed a button on a response pad.

Data Analysis

We investigated the data in two different sets of analyses. The first set focused on visual responses, examining eye-movement measures on the target object from scene onset (i.e.,

including the preview time). Our aim was to address the controversy surrounding effects of saliency and context by reporting eye-movement measures that have also been used in previous research (e.g., Loftus & Mackworth, 1978; Underwood & Foulsham 2006; Underwood et al., 2008; Vo & Henderson, 2009). The second set of analyses focused on the linguistic responses, especially on the impact that visual, linguistic and attentive features have on the act of naming, and on the order of mention.

Search latency, that is the time from scene onset until a fixation lands for the first time on the object of interest, was used. Even if an object naming is not a search task, we align with the terminology of previous literature where such a measure has been used. This measure, as noted, has led to inconsistent results in the literature (e.g., Underwood et al., 2008; Vo & Henderson, 2009). Note that the search is not cued to a specific target in this experiment. Search latencies therefore refer to the object of interest, i.e., the object on which the experimental manipulation was carried out. This definition of search latency is identical to the one adopted by Underwood and Foulsham (2006) in their search task[4] where the manipulation of congruency and saliency was not on the cued target, but on other objects present in the scene.

We also analyze first fixation duration and total gaze duration on the object of interest. We expect salient objects to be fixated longer, for the first time, than non-salient objects, as they carry more low-level information, which can be exploited by language processing

4 Underwood and Foulsham (2006) call this measure time prior to fixation.

for naming. A visually salient object is more informative than a non-salient one, hence more attention is allocated to fully extract the information associated with it. Obviously, this claim holds when the identity of a visually salient object can be recognized, and its linguistic denotation can be retrieved. The implication of this argument is that there might be visually salient regions in the scene that cannot be actually recognized as objects, and therefore will not be viable candidates for naming. However, if incongruent objects are semantically more informative than congruent objects, then they should be looked at overall (total gaze) for a shorter period of time, as they deviate from the overall context their information can more easily memorized. Within the predictive coding framework, an incongruent object is more distant in representational space the than congruent objects. Thus, congruent objects are expected to interfere with one another by attracting attention. This should result in an overall decrease of fixations on the incongruent object. By interference, we mean that congruent objects would receive a similar level of co-activation, which would, in turn, attract attentional resources on them. In practice, we expect less re-fixations to the incongruent object as a result of co-activation of contextually congruent objects. If this logic is correct, then we should observe the same effect in Experiment 2, which employs a memorization task. Notice, this expectation contrasts previous studies that used the out-of-context object paradigm, where incongruent objects are found to be fixated for longer than congruent objects (De Graef, Christiaens, and D'Ydewalle, 1990; Henderson, Weeks & Hollingworth, 1999; Loftus & Mackworth, 1978; Underwood et al., 2008; Vo & Henderson, 2009, 2011).

Secondly, we look at time-course measures and investigate how visual saliency and contextual congruency modulate temporal aspects of object naming. We consider a window of 1000 ms before and after naming divided into 80 intervals of 25 ms each. For each time interval, we calculate the empirical logit of fixations (Barr, 2008):

$$emplog(y) = \log((y + 0.5)/(N - y + 0.5))$$

where y is a fixation to the object of interest (0, 1), and N is the total number of fixations to the other objects in the scene within each interval.

In order to capture the non-linear trend of fixations over time, we adopt the growth curve analysis approach (Mirman, Dixon, & Magnuson, 2008), in which Time is represented through orthogonal polynomials. We chose second order polynomials (Linear and Quadratic), as their associated coefficients can still have a plausible interpretation (interpretation becomes challenging, if not impossible, for polynomials of higher order).

The second analysis focuses on naming measures, with the aim of identifying features that predict object naming. For each object in each trial of the experiment, we code whether the object was mentioned (1 = Mentioned, 0 = Not Mentioned) or looked at (1 = Looked At, 0 = Not Looked At). In case a fixation lands outside of the annotated polygon, we rely on the Euclidean distance from the center of mass of the object. An

object is counted as looked at when the visual angle between the center of mass and the fixation position is smaller than 2 degrees, i.e., the size of the fovea.

Each list of mentioned objects was manually transcribed, and the onset and offset of each word was marked. On average, participants started naming 1588 ± 854 ms after the beep, each word had a duration of 775 ± 226 ms and there was an interval of 883 ± 1750 ms between words. Participants most likely did not promptly begin naming at the beep, as they needed to evaluate which objects they attended to during the preview were worth mentioning. In order to correctly associate eye-movements on objects with the words mentioned, we transcribed the words uttered in a given scene using the labels of the corresponding annotated objects (different participants might have used different words to denote the same object, e.g., *desk*, *table*).

We investigate which visual and linguistic factors predict mentioning and looking during naming. The visual factors we consider are: the saliency of the object (Walther & Koch, 2006, Torralba, et. al., 2006), its area in pixel square, and its contextual fit. Both saliency (Salient, Non-Salient) and congruency (In-Context, Out-of-Context) are coded as categorical variables. The linguistic factors we used were the log-transformed frequency of the word associated with the object (obtained from the CELEX-2 database, Baayen, Piepenbrock, & Gulikers, 1996) and the length of the word being produced when naming the object (in milliseconds).

Since we asked participants to name five objects, we also investigate how the impact of these factors changes over the different instances of naming. Here, we focus on the objects mentioned, and use the position of naming (from 1 to 5) as the dependent measure.

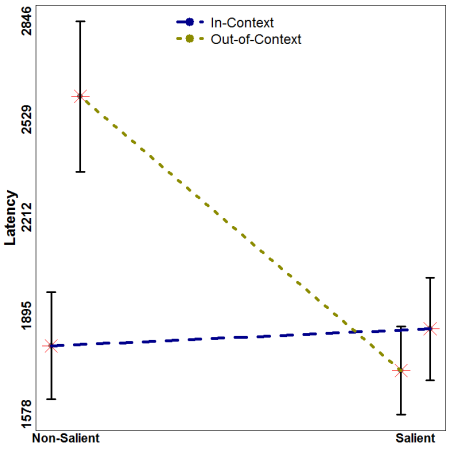
All analyses were performed under the statistical framework of linear mixed effects modeling (LME) as implemented by the R package lme4 (Baayen, Davidson, & Bates, 2008). In LME, the dependent measure is modeled as a linear function of different predictors (fixed effects), and the variance implicit in the multilevel structure of the data is accounted for by grouping based on the random variables of the design. We perform model selection to obtain a minimal mixed-effects model. We compare nested models based on the log-likelihood improvement using chi-square tests.

For example, a model with only a random intercept on participants (depM ~ (1 | Participant); in the syntax of R's lme4 package), is compared to a model with a random intercept also on trials (depM ~ (1 | Participant) + (1 | Trial)). If the log-likelihood of the second model, i.e., the one with the additional parameter (fixed or random) is significantly better than that of the first model, we retain it, otherwise we keep the first model. We start by building the random structure of the model, then we proceed adding fixed effects, e.g., Saliency, and uncorrelated random slopes on it (e.g., 0 + Saliency |

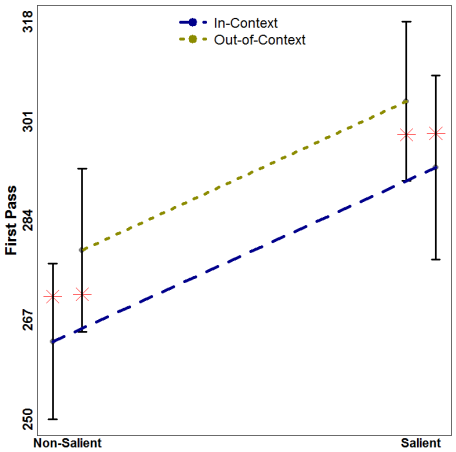
Participant). The inclusion of a random slope accounts for the variability of a fixed effect (here: Saliency) with respect to the grouping level of a random effects (here: Participants). We then include interactions but consider only those which do not violate the subset criterion, i.e., interactions are generated from the subset of main effects included in the model. Factors are included in order of the log-likelihood improvement they bring to the model (the variable which most improves the model fit is included the first, etc.). All factors were centered, i.e., the mean of the factor values across all data points was computed, and then this mean was subtracted from the individual data points. This results, for example, in values of 0.5 and -0.5 for a categorical variable with two factors (or close to these values if there are slight imbalances in the design due to missing values).

We report and discuss the LME model coefficients of the best fitting model. The tables therefore only list those predictors that were retained in the best model. The predictors in the table are ordered following the inclusion order obtained through model selection. Furthermore, for illustrative purposes, we plot the mean of the model fitted values together with the observed mean. If the mean of the model fit falls within the standard error, it means that the model is accurately capturing the data pattern. Note, as the model produces estimates, the mean fit will not always match the observed mean.

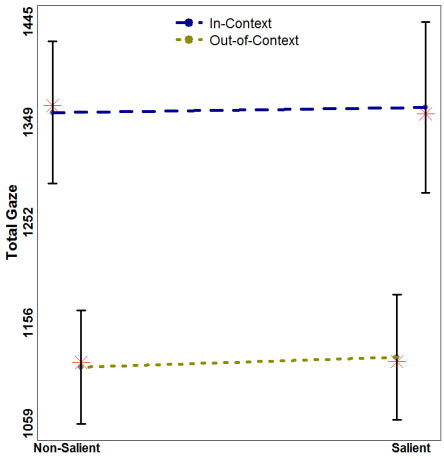
Figure 2. Interaction plots (means and standard error) for different eye-movement measures across experimental conditions in Experiment 1: Saliency (No-Salient, Salient); Context (In-Context; Out-of-Context). Asterisks indicate predicted values according to the LME model.



a) Search Latency: the time spent before landing onto the ROI



b) First fixation: the fixation duration on the ROI during its first inspection



c) Total gaze: Sum of all fixations on ROI

Results and Discussion

Visual Responses. Figure 2 plots the eye-movement measures across the different experimental conditions. Each figure also includes the predicted values for each condition based on the best linear mixed effect model after selection.

In Figure 2(a), we plot search latency, i.e., the time elapsing from the onset of the scene to the first fixation on the object of interest. Here, we find a main effect of Saliency: a

Table 1: Coefficients for the mixed effects model analysis of Mention in Experiment 1. The dependent measures are: search latency, first fixation duration, and total gaze duration. The centered predictors are Saliency (Salient, -0.5, Non-Salient = 0.5) and Context (In-Context = 0.5, Out-of-Context = -0.5)

Search Latency

Predictor	Coefficient
(Intercept)	2007.8***
Saliency	747.7*

Search Latency (Full Model)

Predictor	Coefficient
(Intercept)	2031.9***
Saliency	-437.7*
Context	249.8º
Saliency:Context	-832.2**

First Fixation

Predictor	Coefficient
(Intercept)	285.48**
Saliency	-28.47*

Total Gaze

Predictor	Coefficient
(Intercept)	1227***
Context	240.01***

º $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Salient object is looked at sooner than a Non-Salient one. While the plot seems to indicate an interaction between Saliency and Context, this interaction was not included in the model during model selection. This is because of the way the step-wise forward model selection operates. Since the main effect of Context failed to be included in the model, i.e., it was not significant, any interaction with Context could not be included either to respect the subset criterion (refer to the previous section for details on model selection). We therefore also fitted a full model, which includes Context as a main effect

and an interaction with Saliency. This model (also given in Table 1) shows that incongruent objects are inspected for the first time later, if they are not salient.

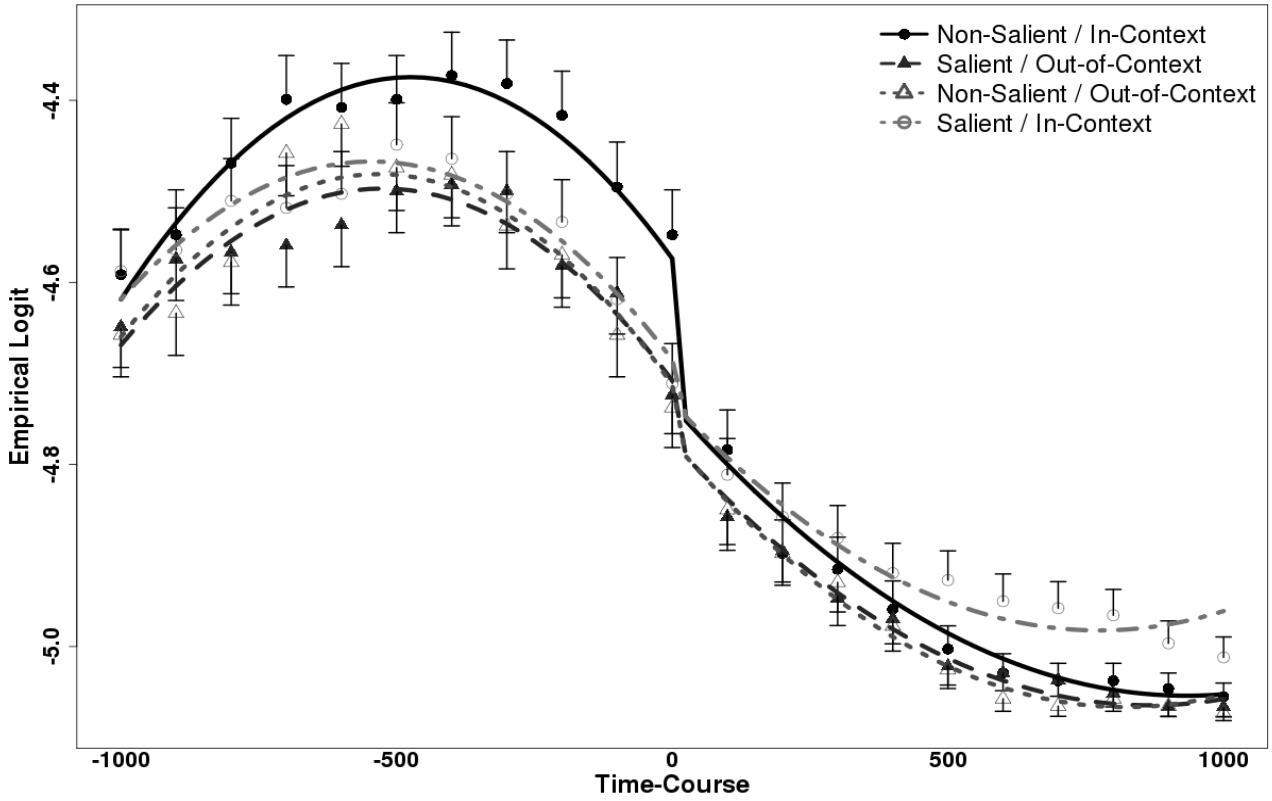
This result indicates that the saliency of objects is actively utilized by the visual system to guide attention in order to select objects that could be interesting for naming. Moreover, the saliency of an object compensates for the delay due to incongruence by boosting its visual appearance. In fact, incongruent objects are inspected for the first time later, i.e., it took longer to identify the target object from scene onset, but only if it is not salient. This result contrasts with previous research showing that saliency effects are overridden by the cognitive evaluation of the scene information (e.g., Einhauser et al., 2008). It also contrasts with previous studies in which incongruency shortened search latencies in a memory task (but not in a search task; Underwood & Foulsham 2006), and in a comparative search task (Underwood, et. al., 2008). Furthermore, an interaction between visual saliency and contextual congruency was not found in either of these previous studies. If an object is recognized as incongruent during initial scene gist processing, such object should be quickly attended. On the contrary, our results show that eye-movements are mostly directed to contextually congruent objects, particularly if they are visually salient. The Cognitive Relevance Framework is consistent with such results, as objects which do not fit the contextual congruency of the scene have longer latencies to first inspection. Specifically, it seems that contextually related objects are more quickly identified (Mack & Eckstein, 2011, Hwang, et at. 2011), hence facilitating also their evaluation as potential naming targets. This result is also compatible with the predictive

coding framework, where the identification of an incongruent object is delayed as a result of the interference between congruent objects competing for attentional resources.

In Figure 2(b), we plot first fixation duration, defined as the time spent on the object of interest during the first fixation. We confirm a main effect of Saliency: salient objects are fixated longer than non-salient objects. A salient object carries more low-level information than a non-salient one, thus visual attention is allocated for a longer period of time to extract relevant information before moving on to the next object; refer to Table 1 for the list of coefficients.

In Figure 2(c), we plot total gaze, which is the sum of all fixations on the object of interest during the whole trial. We find a main effect of Context: out-of-context objects have shorter total gaze duration than In-Context objects. This finding contrasts with previous literature, where incongruous objects have been claimed to attract more attention than congruous objects, as they are contextually more informative. In a cross-modal task such as naming, relevance is evaluated as the product of linguistic and visual factors. Thus an out-of-context object might be visually more informative as it differs from the overall scene context. However, at the same time, an incongruent object is not semantically related to the other objects in the scene, which might make it harder to name. The naming of congruent objects, on the other hand, is boosted by the contextual co-presence of related objects. Contextually congruent objects interfere with each other in

Figure 3. Naming Curve for Experiment 1: time course plot of fixation probability (empirical logit) on the object of interest across conditions before and after naming (from -1000 ms to 1000 ms). The empirical observations are represented as points, while the lines represent LME predicted values.



the representational space. This competition subtracts attentional resources from the incongruent object, which is more distant in representational space than congruent objects. Moreover, more broadly, as an incongruent object is more informative than a congruent object because it does not interfere with the contextually congruent objects, it would require a less intense visual processing to be memorized. If this assumption is

Table 2: Coefficients for the mixed effects model analysis of the empirical logit of fixation probability in Experiment 1. The standardized and centred predictors are Time (Linear, Quadratic; 80 slices of 25 ms each), Region (Before = 0.5; After = -0.5), Saliency (Salient = 0.5, Non-Salient = -0.5); Context (In-Context = 0.5; Out-of-Context = -0.5) and Rank (continuous variable 1–5).

Predictor	Coefficient
(Intercept)	-4.649***
Region	-0.482***
Time-Linear	-3.889***
Time-Quadratic	- 0.181
Rank	0.022
Context	- 0.009
Saliency	0.128
Region:Time-Quadratic	-3.187***
Rank:Time-Quadratic	-0.079***
Region:Context	0.181***
Rank:Saliency	-0.013
Rank:Context	0.002
Region:Rank:Context	-0.058***
Region:Context:Saliency	-0.269***
Rank:Context:Saliency	0.076
Region:Rank:Saliency	-0.035***

º $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

correct, we should see an identical effect of congruency also during memorization (Experiment 2).

Finally, in Figure 3, we graph fixation probability on the object of interest (empirical logit), before and after naming, across the four experimental conditions. We also include a numeric variable coding the position in which the naming occurred (Rank: 1-5) to show how the impact of visual saliency and contextual congruency changes with different instances of naming.

We will discuss the coefficients of those predictors that were significant after model selection, as tabulated in Table 2. Positive coefficients indicate an increasing probability of fixating the target object, whereas negative coefficients indicate a decreasing probability. Note that for an interaction with multiple terms, the sign of the coefficient (positive vs. negative) is obtained by multiplying the coefficient of the interaction with the values of the factors that participate in the interaction (recall that factors are centered, i.e., take on values of +/-0.5). For example, if we are interested in whether a Congruent object (Context = 0.5) was fixated more or less Before being named (Region = 0.5), then we multiply the coefficient for that interaction with the values of the factors we are interested in (i.e., 0.181 * 0.5 * 0.5).

We find main effects of Region (Before naming has less fixation than After), and Time-Linear, with fixation probability decreasing over time. Given our experimental design, important effects will show up as interactions, rather than main effects. There is a positive interaction of Region and Context, with Out-of-Context objects fixated less often Before being named (two-way interaction Region:Context), especially when the object is named at a higher Rank (three-way interaction Region:Rank:Context). We also find an interaction of Region and Time-Quadratic: fixation probability first increases, and then falls again before naming, hence the quadratic slope. A quadratic slope is found also with Rank (two-ways interaction, Rank:Time-Quadratic), which indicates that fixation probability falls after naming more rapidly for objects named at later ranks. Possibly,

visual attention is shifted more quickly to other naming targets, as they are uttered, to optimize the the allocation of attentional resources.

Turning to Saliency and Context, we find that Salient objects are fixated more often if they are In-Context, especially Before being named (three-way interaction Region:Context:Saliency). Moreover, there are fewer fixations on salient objects when an object is named at later ranks, and this effect is especially prominent in the region Before naming (three-way interaction Region:Rank:Saliency). As we shall see in the analysis of features predictive of naming, salient objects are named at earlier ranks (see Table 4). This means that the probability of fixating a salient object decreases across different instances of naming, as it becomes less likely that it will be named.

When comparing the results on visual responses with Underwoord and Foulsham's (2006) study, where saliency and congruency were investigated in a memory and search task, we find some similarities but also many differences. The authors found that during memorization a visually salient object is fixated earlier (search latency), but not longer than the less visually salient object. Moreover, incongruent objects were fixated earlier when non-salient, and fixated for longer than congruent objects (first fixation), regardless of their saliency. During search, saliency had no influence on fixation, and the only positive result was an interaction, whereby a salient object was fixated earlier when an incongruent object was present in the scene. They interpret this interaction as evidence

that incongruent objects are rapidly detected during the processing of gist, and visual attention is oriented to the most conspicuous object as a reaction to it. Note, however, that this interaction was indirect, as Underwood and Foulsham (2006) did not systematically manipulate visual saliency and contextual congruency on the same target object. We found earlier search latency for salient object in line with the result of their memory task. However, we found that initial fixations on salient objects are longer, rather than shorter, compared to non-salient objects. We argued that saliency is used to evaluate whether a target is worth being mentioned. We find that incongruent objects are looked at later, and less frequently than congruent objects, contrasting with Underwood and Foulsham (2006) and Underwood, et. al. (2008). We argued that an incongruent object might be visually more informative, but at the same time its lack of context might make it harder to name. Differently from both studies, we find an interaction between visual saliency and contextual congruency, whereby the likelihood of attending an incongruent object increases when such object is visually salient. This indicates that both types of information are concurrently evaluated when guiding visual attention during a naming task.

Naming objects requires to access and use visual information differently than both memorization and search. Similar to memorization, in the naming task targets are not cued, but at the same time, a pool of search targets has to be selected for naming.

In the naming task, salient objects are likely to be selected as targets (as no object is cued), but at the same time contextual information of the scene becomes important to finalize the selection of targets from the pool of possible objects.

In the next section, we will see that the way in which bottom-up and top-down information is visually processed has direct consequences for which objects are selected for naming, and for their order of mention.

Linguistic Responses. As mentioned in the Data Analysis section, each annotated object in each trial was coded as mentioned (0, 1) or looked at (0, 1). We find a small number of cases where an object is mentioned but not looked at (1.8%). It seems that some objects are recognized parafoveally, even though we tried to account for this when coding the data. An analysis of the eye-tracking data indicates that when an object was named but not fixated, the nearest fixation fell with 4.1 ± 2.5º visual angle from the target centroid. The highest percentage of objects was both looked at and mentioned (35.9%). There was almost an equal number of objects that were not looked at and not mentioned (30.8%) and looked at and not mentioned (31.5%).

As a next step, we determined which visual and linguistic features are predictive of naming, regardless of whether the named object is the object of interest, a competitor, or another object in the scene. We predicted Mention (0, 1) as a function of visual,

Table 3: Coefficients for the mixed effects model analysis of Mention in Experiment 1. The standardized and centered predictors are the size of object (Area), the duration of the gaze on the object (Gaze), the log frequency of the word uttered (LogFrequency), the mean Saliency of the object, and the probability of the the object being fixated during preview (Preview).

Predictor	Coefficient
(Intercept)	-0.086**
Gaze	-1.156***
Area	-0.511***
Saliency	0.580***
LogFrequency	-0.173*
Preview	0.702***
Area:LogFrequency	-3.068***
Area:Saliency	3.073***
Area:Preview	-1.252*
Saliency:LogFrequency	0.603**
Gaze:Saliency	-1.104**
Area:Saliency:Preview	29.687**
Gaze:Saliency:Preview	-15.461*

º p < 0.10, * p < 0.05, ** p < 0.01, *** p < 0.001

linguistic, and attentional factors in a linear mixed effect model. The visual factors were saliency of the objects, its area[5], and its contextual fit.

The linguistic factor was the frequency of the name of the object as estimated from the CELEX database. We also included as control variables the gaze duration on the object and the probability of looking at the object during preview (see Data Analysis for details on how the factors were computed).

Table 3 lists the coefficients of the resulting mixed model. Positive coefficients here indicate a higher probability of naming that object, whereas negative coefficients indicate

5 Area was included also in previous analysis, but excluded during model selection.

a lower probability. We find that gaze duration and saliency are positive predictors of naming, while the area of the object and the log-frequency of the word used for naming are negative predictors (main effects: Gaze, Saliency, Area and LogFrequency). Moreover, having looked at the object during the preview time increases the probability of naming it (main effect: Preview). When looking at the interactions, we find that large objects with high saliency are named more often (two-ways interaction Area:Saliency), especially when the object was previewed (three-way interaction Area:Saliency:Preview). Longer gazes on salient objects are predictive of naming only if the object has been previewed (three-ways interaction Gaze:Saliency:Preview). In fact, a salient object, which has longly being fixated, it is actually less likely to be named (two-ways interaction Gaze:Saliency). Finally, a lexically frequent object is more likely to be mentioned if it is salient (two-ways interaction Saliency:LogFrequency), but not if it is large (two-ways interaction Area:LogFrequency).

The first implication of these results is that the saliency affects positively the likelihood of naming an object: a finding that challenges purely top-down approaches to visual cognition during goal-oriented tasks (e.g., Einhauser et al., 2008; Henderson et al., 2009). In contrast to the psycholinguistic finding that lexical access, and consequent naming, is boosted by lexical frequency (Almeida et. al. 2007), we find that frequent objects are less likely to be named. Presumably, many frequent objects (e.g., STREET) are not visually interesting; this is confirmed by the negative interaction with area. Lexical frequency,

Table 4: Coefficients for the mixed effects model analysis of Rank (1–5) in Experiment 1. The standardized and centered predictors are the size of object (Area), the duration of the word named (WordDuration), the log corpus frequency of the word uttered (LogFrequency), the mean Saliency of the object, and the probability of having looked at the object during preview (Preview).

Predictor	Coefficient
(Intercept)	2.245***
WordDuration	0.056
Area	0.983
LogFrequency	-0.047
Preview	-3.415***
Saliency	-0.221**
Area:LogFrequency	11.693**
Saliency:LogFrequency	1.041º
WordDuration:Area	14.670*

º $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

however, is also modulated by saliency: a lexically frequent object becomes more interesting if it is also salient. This provides evidence for an interaction between linguistic and low-level visual information, and shows that lexical access is not independent of perceptual and conceptual variables involved in the context (but see Almeida et al., 2007). In addition to this, our data also confirms that visual information attended to at the early stages of the trial is important for predicting naming: objects targeted during the preview are likely to be named. This effect is more pronounced if such objects are visually salient and have a large area.

Naming objects is a task that demands the integration of different sources of information. However, the influence of these different sources, and the way in which they are accessed, could change over the course of the trial. It is possible that factors guiding the naming at earlier ranks no longer exert influence at later ranks. We therefore consider

only those cases in which an object was named, and use Rank, i.e., the position on which naming occurred, as the dependent measure. For this analysis we also included as a covariate the duration of the word being uttered. Here, positive coefficients refer to later instances of naming, whereas negative coefficients to earlier instances.

We found negative effects of Preview and Saliency: previewed objects are less likely to be named at later stages of naming (see Table 4). This indicates that the probability of previewing an object is predictive only during early naming, e.g., first or second object named. In the same vein, salient objects are named at earlier rather than later ranks. We still found a marginal interaction between saliency and word frequency (two-ways interaction Saliency:LogFrequency), indicating that more frequent and highly salient objects are more likely to be produced even at later ranks. However, objects named later tend to have a large area and are associated with highly frequent words (two-ways interaction Area:LogFrequency). Once the most visually and linguistically interesting objects have been named, language processing is directed to large background objects to continue the naming task. It seems clear that the way cross-modal information is accessed changes across naming instances. In particular, the effect of saliency on naming tends to decay, but probably this result comes about because there are fewer highly salient objects in the scene to be named. A similar reasoning can be applied to preview probability. The objects with the highest probability during the preview are spelled out as soon as naming begins. After this, the visual system turns to previously unattended objects to keep

sourcing material for the ongoing process of naming. Finally, we also observed an interaction between area and word-duration, whereby object with a large area, which are also associated to long words, are produced only at later ranks (two-ways interaction WordDuration:Area). In order to optimize naming efficiency, participants spell-out background objects which have short names earlier, before resorting to background objects, which are associated with longer names.

We failed to find a significant effect of contextual congruency in the analysis of either the mention or the rank data. This could be because both analyses included all objects, rather than just the target objects.[6] We therefore ran additional analyses of both the mention and the rank data but included only the target objects. The results confirm that contextual congruency does not play any role in the naming patterns, i.e., the contextual variable is not included as significant during model selection. We conclude that while visual attention focuses on congruent objects (as our analysis of fixation latencies and duration showed), congruent and incongruent objects have the same likelihood of being selected as naming candidates.

To summarize our results, we find that visual saliency had a clear effect during a naming task, both on the likelihood of mentioning a certain object and on the probability of fixating it. Salient objects are mentioned earlier than non-salient ones, suggesting that the

6 There is only one out-of-context object that could be mentioned for every five naming instances, i.e., there are more congruent objects that can be named compared to incongruent once. Thus, the incongruent condition is unbalanced when considering the full dataset.

more objects that have been named, the smaller the effect of saliency becomes. An alternative explanation is that the selection of objects to be named demands that visual attention is allocated to non-salient objects once it has used up the salient ones. Note that the effect of saliency we found is not predicted by theories of visual attention that assume strong top-down control during goal-directed tasks. When looking at the effect of context, we found that incongruent objects are looked at less often than congruent objects. Fixations on incongruent objects are boosted only when the objects are visually salient. This indicates that fixations during naming are guided by a combination of bottom-up and top-down information. However, when looking at the linguistic pattern of naming, we find that only bottom-up information seems to play a role. In fact, an incongruent object has the same likelihood of mention as a congruent one.

The present experiment employed a novel task, viz, object naming. While this task has conceptual advantages over purely visual tasks (the cross-modal nature of naming is likely to encourage the interaction of both top-down and bottom-up processes, as we argued in the Introduction), it has the disadvantage of not being directly comparable to results in the existing literature, which were obtained using more standard tasks like visual search or memorization. This makes it hard to ascertain whether the results we found in Experiment 1 were task-specific, or generalize to other tasks. Therefore, in a follow-up experiment, we looked at eye-movement responses during a memorization task using the same set of stimuli and experimental conditions. A comparison of visual

responses between the two tasks makes it possible to disentangle the influence of language processing on visual attention from effects exerted by saliency and congruency in non-linguistic tasks such as memorization.

Experiment 2

We tested how the saliency and the contextual congruency of an object influence eye-movement responses during the memorization of scenes in preparation for a recall task. In essence, the logic of this experiment is similar to the studies of Underwood and Foulsham (2006) and Henderson, et. al. (1999), where memory was tested in preparation for a recognition test and no particular object was cued as being of special importance to the participants.

These studies have shown that salient objects are looked at faster than non-salient objects (Underwood & Foulsham, 2006), and that total gaze duration on incongruent objects is longer than on congruent objects (Henderson, et. al., 1999). Both studies did not find an interaction between saliency and congruency.

Method and Data Analysis

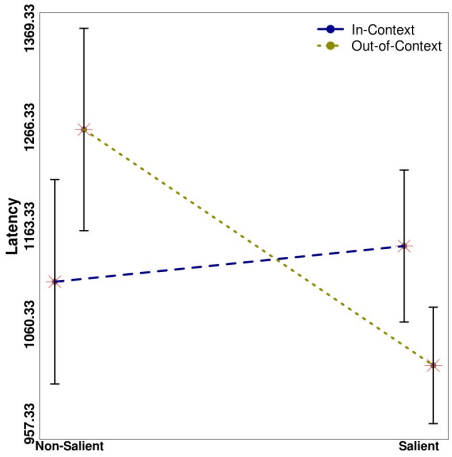
Twenty-four native speakers of English, all students of the University of Edinburgh, were asked to preview each scene for five seconds, after which it was removed, and they had to

recall as many objects as possible. Their recall responses were recorded using a lapel microphone. When participants could remember no further objects for a given scene, they had to press a button on the keypad to go to the next trial. The task was explained using written instructions, and the participants were paid five pounds for their participation.
The experiment employed the same design as Experiment 1, crossing the factors Saliency and Congruency. It also re-used the scenes of Experiment 1, which contained an object of interest whose saliency and congruency were manipulated. The presentation of trials followed the same procedure as Experiment 1, and we used the same apparatus to monitor eye-movement responses.

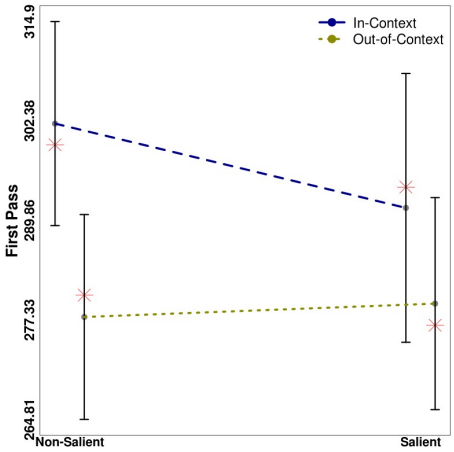
The analysis will exclusively focus on the eye-movement measures on the target object during the five second preview. Thus, we will not include in the analysis any eye-movement data recorded during recall phase. In particular, we analyze search latency, that is the time from scene onset until a fixation lands for the first time on the object of interest, first fixation duration and total gaze duration on the object of interest.
We analyze our dependent measures using linear-mixed effects models, as explained in Experiment 1. As we are interested in the interaction between saliency and congruency, we will report directly only full-models containing both the predictors Saliency and Context as main effects and in interaction[7].

7 We also used model selection and did not found any significant interaction, in any of the measures investigated.

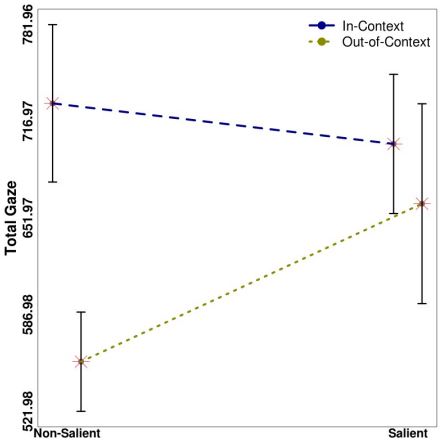
Figure 4. Interaction plots (means and standard errors) for different eye-movement measures across experimental conditions in Experiment 2: Saliency (No-Salient, Salient); Context (In-Context; Out-of-Context). Asterisks indicate predicted values according to the LME model.



a) Search Latency: the time spent before landing onto the ROI



b) First fixation: the fixation duration on the ROI during its first inspection



c) Total Gaze: Sum of all fixations on ROI

Results and Discussion

Figure 4 plots means and standard errors of the eye-movement measures on the target object across experimental conditions, and includes the predicted values of the linear mixed-effects models.

On search latency plotted in Figure 4(a), we find similar trends to what we saw in Experiment 1 (refer to Figure 1); however, none of the predictors reach significance.

Table 5: Coefficients for the mixed effects model analysis of different eye-movement measures in Experiment 2. The dependent measures are: search latency, first fixation duration, and total gaze duration. The centered predictors are Saliency (Salient, -0.5, Non-Salient = 0.5) and Context (In-Context = 0.5, Out-of-Context = -0.5)

Search Latency (Full Model)

Predictor	Coefficient
(Intercept)	1082.63***
Saliency	23.19
Context	-16.37
Saliency:Context	-205.94º

First Fixation (Full Model)

Predictor	Coefficient
(Intercept)	286.65***
Saliency	2.89
Context	20.39
Saliency:Context	7.861

Total Gaze (Full Model)

Predictor	Coefficient
(Intercept)	564.25***
Saliency	-41.24
Context	63.59*
Saliency:Context	33.18º

º $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Participants tend to identify salient objects more rapidly, especially when they are incongruent (marginal two-ways interaction: Saliency and Context). Incongruent objects are not identified any faster than congruent objects.

When looking at first fixation duration plotted in Figure 4(b), there is a trend of longer first looks to congruent objects, which appears to be reduced when such object is also salient. However, none of the predictors reach significance.

Finally, when looking at the total gaze duration plotted in Figure 4(c), we find a significant main effect of Context, whereby incongruent objects are looked at less than congruent objects; this effect is partially compensated for by saliency, as indicated by the marginal interaction in Table 5. This finding corroborates what we observed in Experiment 1, and contrasts with the study of Henderson, et. al. (1999), where incongruent objects were looked at longer than congruent objects in a memorization task. An explanation for our finding is that incongruent objects are more informative than congruent objects and thus need less attention in order to be memorized. From a predictive coding perspective, an incongruent object is semantically more distant in representational space than congruent objects. This implies that congruent objects interfere with each other on the representational space, decreasing their likelihood to be accurately remembered. An incongruent object instead escapes the interference in such representational space, as it triggers a prediction error, and this makes it easier to remember. In line with Underwood and Foulsham (2006), saliency seems to attract earlier and longer first fixations to the object of interest (though this tendency did not reach statistical significance).

To summarize, we observed very similar patterns in memorization and object naming, in all three measure we looked at. However, we found a significant effect of Context only for total gaze duration. This result is quite important, as it contrasts from previous

research on the topic, and shows that incongruent objects are attended overall less than congruent objects, both during object naming and memorization. This finding might indicate that the informativeness of an incongruent object makes the object easier to remember, which in turn implies a less extensive visual processing. More attention is required to distinguish, and remember, semantically related objects, as they are closer in the representational space than an object deviating from the common context. More research is needed to more specifically test whether the memorability of an object increases or decreases according to its contextual fit, and whether the effect varies for cueing versus non-cueing paradigms. Moreover, the memory for verbal recall task might trigger similar inspection strategies as a purely object naming task, i.e., attend as many objects as possible to recall/name them later. Perhaps a comparison of object naming with a visual search task would help highlighting the actual difference between purely visual and linguistically mediated tasks. Both search latency and first fixation duration, instead, were not significantly mediated by the predictors Saliency and Context. This fact might be caused by the time pressure under which the participants performed the memorization task. As the preview time was restricted to five seconds, participants presumably tried to scan as many objects as possible before the scene was removed. The pressure to look at a wide range of objects in preparation for recall might have increased the variance in the search latency and first fixation duration measures for the objects of interest. As search latency and first fixation are both measures of first object identification, a pressure of moving onto other targets might have influenced the

participants on the time available to evaluate what object should have been look first, i.e., search latency, and for how long, i.e., first fixation duration.

General Discussion

Scene understanding requires access to both low-level (stimulus-based) and high-level (knowledge-based) visual information. However, there are conflicting views regarding when these two types of information are utilized to guide visual attention. A bottom-up theory of visual processing assumes that low-level features guide visual attention (e.g., Itti & Koch, 2000). Top-down approaches, such as the Cognitive Relevance Framework (Henderson et al., 2009), in contrast, posit that high-level contextual information is the main source of attentional control during scene understanding, leaving low-level information to play only a minimal role for a given task. It is likely that the divergent results are due to the different experimental tasks used, and especially one crucial factor seems to be the presence of an explicit target object: A search task involves a target object that has been cued ahead of the trial, while a memorization task does not. It seems likely that setting a target changes the way in which context is processed, as visual attention is then more susceptible to the semantic content of the scene, such as the co-occurrence of objects (e.g., a BALL is usually found on the FLOOR). Even though Underwood et al. (2008) tried to address this issue by manipulating visual saliency and contextual congruency in a comparative search task, the results remain inconclusive as

their task is highly structured, and it seems likely that participants develop specialized scanning strategies (Underwood, 2009).

In the present study, we examined the interplay between stimulus-based and knowledge-based information on attentional guidance, and linguistic performance, during an object naming task, and compared the eye-movement responses in such a task with those in a memorization task. We hypothesized that visual attention and language processing have to share information across modalities to achieve object naming. Thus, an interaction between low- and high-level information of the visual scene is expected to take place and manifest itself both in visual and linguistic responses. Naming objects is a simple language production task for which a large experimental literature exists in psycholinguistics. It is clear from that literature that linguistic responses are mediated by contextual constraints, both linguistic and non-linguistic (e.g., Bartram, 1974; Snodgrass, 1980; Meyer et al., 1998; Griffin & Bock, 1998; Hocking et al., 2009). Moreover, unlike search, a naming task does not explicitly set a specific target object beforehand, but requires participants to perform an implicit selection of targets according to the visual and linguistic information available. Thus naming is a task that puts visual saliency and contextual congruency directly in competition, making it ideal for investigating how these two types of information are processed when choosing objects to name. Moreover, the task also makes it possible to examine how linguistic information (such as word frequency) interacts with saliency and context during the process of naming. If object

naming is performed through a cross-modal processing of various sources of information, we expected interactions to emerge and support target-selection with the goal of mention.

When analyzing the visual responses, we found that search latency was longer for incongruent objects. In line with the Cognitive Relevance Framework, an incongruent object is not readily inspected after scene onset as it does not fit the cognitive top-down constraints imposed by the task goals. Selecting objects to name, in fact, demands a contextual evaluation of the scene (e.g., a kitchen scene is likely to contain a SPOON and a PLATE), which is known to be involved in such tasks (Damian, et. al. 2002). This effect is also compatible within the predictive coding framework (e.g., Hinton 2007), according to which contextually related objects share representational space attracting attention overt them, and consequently delaying the identification of the incongruent object. However, the effect of context disappears when the incongruent object is visually salient. This finding is consistent with what was observed by Underwood and Foulsham (2006), who found that the presence of an incongruent object in the scene boosted looks to a salient object also in the scene. However, our finding also significantly differs from Underwood and Foulsham (2006), as in our study, the manipulation of congruency and saliency were applied to the same target object and not to separate non-target objects. This made it possible to evaluate how the joint contribution of congruency and saliency properties of an object modulates visual responses.

Importantly, our results contrast with a strong interpretation of top-down guidance during goal-oriented tasks, which holds that visual saliency is not expected to play a role. Moreover, when looking at first fixation duration, we find that salient objects are looked at longer than non-salient ones. This result suggests that the visual saliency of an object enhances its likelihood of being looked at and subsequently selected as a linguistic target to be named. This result also differed from that of Underwood and Foulsham (2006), who did not find any effect of saliency on the first fixation. Finally, we also find that incongruent targets are overall (i.e., in total fixation duration) looked at significantly less than congruent targets, which is at variance with the idea that incongruency increases visual informativeness. Rather, in a naming task, an incongruent object is less important because it is linguistically irrelevant given the scene context. Moreover, an incongruent object might be also more memorable, as this effect of contextual incongruency was replicated in the follow-up memorization study, discussed in more detail below. The memorability of an incongruent target might be a consequence of the fact that, as it is deviantit does not interfere in representational space with congruent objects, hence making it more likely to be remembered.

When analyzing the linguistic responses, we found that the visual saliency of the object of interest affects naming patterns. Saliency was a significant predictor of whether an object is named or not, with salient objects more likely to be named. Saliency had its greatest effect when naming begins, i.e., after a 1500 ms preview. An analysis of the

naming sequence revealed that salient objects were also named earlier than non-salient objects. Crucially, however, we did not find contextual congruency to play a role in determining whether an object is named, or at which position in the naming sequence it occurs. Incongruent objects had the same likelihood to be named as congruent ones. While visual attention focuses on congruent objects during naming (as evidenced by our analysis of search latencies and fixation durations), the language processor seems to show no preference for congruent objects as potential naming candidates.

These findings allow us to construct a timeline of naming: language processing sources salient objects from the visual system first, which makes them naming targets to prioritize. This interpretation is compatible with low-level information being processed during early stage visual processing. Then, visual attention is allocated mainly to contextually relevant objects, and further naming targets are selected from this pool. This result stems from the fact that incongruent objects are looked at later because they are inconsistent with the context of the scene, and hence they tend to be excluded from the naming pool. Thus, the remaining targets selected to finalize the naming task are predominantly non-salient, often background objects (large area). At this stage, contextual congruency no longer plays a role, and congruent and incongruent objects have the same likelihood of being selected for naming.

The aim of the naming experiment reported in this paper was to investigate how context and saliency interact to guide visual attention. The use of a linguistic task, however, enabled us to also determine the impact of linguistic factors on naming. We found that objects with more frequent names are less likely to be named; furthermore, frequency interacts with the visual properties of the objects being named. In particular, we found that salient or large objects that are lexically frequent are more likely to be named. From a linguistic perspective, this result suggests that the lexical frequency of the word used to refer to the visual object is not sufficient to decide whether an object has to be named. Rather, it seems that lexical frequency has to interact with other properties of the object, such as its visual saliency, to make it a viable naming candidate. It is a standard result in psycholinguistics that frequent words are retrieved more quickly during lexical decision and naming. However, our results indicate that speed of retrieval is not the main determinant for an object being selected for naming. Rather, such selection is accomplished by combining linguistic information with both low- and high-level visual information.

In order to test whether the effects we observed in our naming experiment were particular to naming (Experiment 1), we also looked at eye-movement responses during a memorization task using the same experimental materials (Experiment 2). The memorization task also ensures comparability with previous work on saliency and object congruency, some of which has used memorization (Underwood & Foulsham, 2006,

Henderson, et. al., 1999). Experiment 2 found trends comparable with the object naming study, but failed to achieve statistical significance for search latency and first fixation duration. In particular, there was a trend of looking at salient objects earlier and longer for the first time, in line with results by Underwood and Foulsham (2006). This result suggests that salient information tends to be activated at earlier stages of visual processing. We argued that the failure to find significant effects in these measures might be due to the task demands: Since participants had a fixed preview time of five seconds, which put them under time pressure to memorize as many objects as possible. Such time pressure (which was not present in the naming study, where the scene remained visible until the participants ended the trial) might have increased the variance of first pass measures, i.e., search latency and first fixation duration, on the object of interest. The rationale behind this argument is that search latency and first fixation are both measures of initial object identification. So, in a task demanding scenes to be widely inspected and memorized for object recall, participants did not have enough time to evaluate which object should have been looked first (search latency) and how much should have been attended (first pass). However, on total gaze duration, we replicated what we found during object naming: incongruent objects were fixated overall less than congruent object. This finding contrasts with the study of Henderson, et al. (1999), which found the opposite effect,. This could suggest that incongruent objects, as they deviate from the scene context, are more memorable than congruent objects and hence need to be attended less. More research is needed to elucidate the effect of congruency on memorability of

objects. Overall, we did not find any significant interactions between saliency and congruency in Experiment 2, which might indicate that during a purely visual task, these two types of information are selectively used, rather than interactively, as Experiment 1 demonstrated for naming.

Understanding which factors determine an object's importance in a scene has significant implications beyond visual cognition. In computer vision, for instance, the accuracy of object detectors or automatic image annotation could be improved by accurately evaluating the visual and linguistic features involved. In future work, we plan to utilize the insights gained in the present work to design models and algorithms that are able to integrate visual and linguistic information in a similar fashion to what humans do when they perform naming tasks.

References

Almeida,J., Knobel, M., Finkbeiner,M., Caramazza, A (2007). The locus of the frequency effect in picture naming: When recognizing is not enough. Psychonomic Bulletin & Review, 14 (6), 1177-1182

Allison, B., Keller, F., & Coco, M. I. (2012). A bayesian model of the effect of object context on visual attention. In N. Miyake, D. Peebles, & R. Cooper (Eds.), Proceedings of the 34th annual conference of the Cognitive Science Society. Sapporo.

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.

Baayen, R., Piepenbrock, R., & Gulikers, L. (1996). Celex2 [Computer software manual]. Linguistic Data Consortium, Philadelphia.Baluch, F., & Itti, L. (2011). Mechanisms of top-down attention. Trends in Neurosciences, 34, 210-224.

Barr, D. (2008). Analyzing visual world eye-tracking data using multilevel logistic regression. Journal of memory and language, 59(4), 457-474.

Bartram, D. (1974). The role of visual and semantic codes in object naming. Cognitive Psychology, 6, 325-356.

Bonitz, V. S., & Gordon, R. D. (2008). Attention to smoking-relate and incongruous objects during scene viewing. Acta Psychologica, 129, 255-263.

Borji, A., Sihite, D. N., & Itti, L. (2012). Quantitative analysis of human-model agreement in visual saliency modeling: a comparative study. IEEE Transactions on Image Processing, 1-16.

Brooks, D. I., Rasmussen, I., & Hollingworth, A. (2010). The nesting of search contexts within natural scenes: evidence from contextual cuing. Journal of Experimental Psychology: Human Perception and Performance, 36(6), 1406-1418.

Castelhano, M., & Heaven, C. (2010). The relative contribution of scene context and target features to visual search in real-world scenes. Attention, Perception, & Psychophysics., 72(5), 1283-1297.

Clark, Andy. "Whatever next? Predictive brains, situated agents, and the future of cognitive science." *Behavioral Brain Science* (2012): 1-86.

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

Damian, M., Vigliocco, G., & Levelt, W. (2001). Effects of semantic context in the naming of pictures and words. Cognition, 81, B77-B86.

Davenport, J., & Potter, M. (2004). Scene consistency in object and background perception. Psychological Science, 15, 559-564.

De Graef, P., Christiaens, D., & Ydewalle, G. (1990). Perceptual effects of scene context on object identification. Psychological Research, 52, 317-329.

Di Lollo, V., Kawahara, J., Zuvic, S. M., & Visser, T. A. (2001). The preattentive emperor has no
clothes: a dynamic redressing. Journal of Experimental Psychology. General, 130(3), 479–492.

Eckstein, M., Drescher, B., & Shimozaki, S. (2006). Attentional cues in real scenes, saccadic targeting and Bayesian priors. Psychological Science, 17, 973-980.

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

Einhauser, W., Rutishauser, U., & Koch, C. (2008). Task-demands can immediately reverse the effects of sensory-driven saliency in complex visual stimuli. Journal of Vision, 8, 1-19.

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

Enns, J. T., & Lleras, A. (2008). What's next? New evidence for prediction in human vision. Trends in Cognitive Sciences, 12(9), 327–333.

Evans, K.,K. and Treisman, A. (2005). Perception of Objects in Natural Scenes: Is It Really Attention Free? Journal of Experimental Psychology: Human Perception and Performance,31 (6), 1476–1492.

Foulsham, T., & Underwood, G. (2011). If saliency affects search then why? Evidence from normal and gaze-contingent search tasks in natural scenes. Cognitive Computation, 3, 48-63.

Foxe, J. J., & Simpson, G. V. (2002). Flow of activation from V1 to frontal cortex in humans – A framework for defining "early" visual processing. Experimental Brain Research, 142(1), 139–150.

Friston, K. (2010) The free-energy principle: A unified brain theory? Nature Reviews Neurosciences 11(2):127–38.

Gajewski, D. A., & Henderson, J. M. (2005). Minimal use of working memory in a scene comparison task. *Visual Cognition: Special Issue on Real-World Scene Perception*, 12, 979-1002.

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. (1998). Constraint, word frequency, and the relationship between lexical processing levels in spoken word production. Journal of Memory and Language, 38, 313-338.

Griffin, Z., & Oppenheimer, D. (2006). Speakers gaze at objects while preparing intentionally inaccurate labels for them. Journal of Experimental Psychology: Learning, Memory and Cognition, 32, 943-948.

Hayhoe, M., & Ballard, D. (2005). Eye movements in natural behavior. Trends in Cognitive Science, 9,188-194.

Henderson, J., Brockmole, J., Castelhano, M., & Mack, M. (2007). Visual saliency does not account for eye-movements during visual search in real-world scenes. In R. van Gompel & M. Fisher & W. Murray &R. Hill (Eds.), eye movement research: insights into mind and brain (pp. 538-562). Elsevier.

Henderson, J., Malcolm, G., & Schandl, C. (2009). Searching in the dark: Cognitive relevance drives attention in real-world scenes. Psychonomic Bulletin & Review, 16, 850-856.

Henderson, J., Weeks, P., & Hollingworth, A. (1999). The effects of semantic consistency on eye movements during complex scene viewing. Journal of Experimental Psychology: Human Perception and Performance, 25, 210-228.

Hinton, G. E. (2007a) Learning multiple layers of representation. Trends in Cognitive Sciences 11:428–34

Hocking, J., McMahon, K., & Zubicaray, G. de. (2009). Semantic context and visual feature effects in object naming: An fmri study using arterial spin labeling. Journal of Cognitive Neuroscience, 21, 1571-1583.

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. D., Wang, H-C., Pomplun, M., Semantic guidance of eye movements in real-world scenes. Vision Research, 51(10), 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.

Levelt, W.J.M., Schriefers, H., Vorberg, D., Meyer, A., Pechmann, T., & Havinga, J. (1991). The time course of lexical access in speech production: A study of picture naming. Psychological Review, 98(1), 122-142.

Loftus, G., & Mackworth, N. (1978). Cognitive determinants of fixation location during picture viewing. Journal of Experimental Psychology: Human Perception and Performance, 4, 565-572.

Mack, S. C., & Eckstein, M..P. (2011). Object co-occurrence serves as a contextual cue to guide and facilitate visual search in a natural viewing environment. Journal of Vision, 9(11) (9), 1-13.

Malcolm, G. L., & Henderson, J. (2009). The effects of target template specificity on visual search in real-world scenes. Journal of Vision, 9(11)(8), 1-13.

Malcolm, G. L., & Henderson, J. (2010). Combining top-down processes to guide eye movements during real-world scene search. Journal of Vision, 10(2)(4), 1-11.

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

Mirman, D., Dixon, J., & Magnuson, J. (2008). Statistical and computational models of the visual world paradigm: Growth curves and individual differences. Journal of Memory and Language, 59(4), 475-494.

Parkhurst, D., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention.
Vision Research, 42(1), 107-123.

Potter, M. (1975). Meaning in visual search. Science, 187, 965-966.

Potter, M., Kroll, J., Yachzel, B., Carpenter, E., & Sherman, J. (1986). Pictures in sentences: Understanding without words. Journal of Experimental Psychology: General, 115, 281-294.

Rao, R. & Ballard, D. (1999) Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects, Nature Neuroscience2(1):79.

Russell, B. C., Torralba, A., Murphy, K. P., & Freeman, W. T. (2008). LabelMe: A database and web-based tool for image annotation. International Journal of Computer Vision, 77(1-3), 157-173.

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

Tatler, B. W. (2007) The central fixation bias in scene viewing: selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, 7(14):4, 1-17

Tatler, B. W., & Vincent, B. T. (2008). Systematic tendencies in scene viewing. *Journal of Eye Movement Research*, 2(2), 5: 1-18.

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

Torralba, A., Oliva, A., Castelhano, M., & Henderson, J. (2006). Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. Psychological review, 4(113), 766-786.

Underwood, G. (2009). Cognitive processes in eye guidance: Algorithms for attention and image processing. Cognitive Computation, 1, 64-76.

Underwood, G., & Foulsham, T. (2006). Visual saliency and semantic incongruency influence eye-movements when inspecting pictures. Quarterly Journal of Experimental Psychology, 59, 1931-1949.

Underwood, G., Humphrey, L., & Cross, E. (2007). Congruency, saliency and gist inspection of objects in natural scenes. In R. van Gompel & M. Fisher & W. Murray & R. Hill (Eds.), Eye movement research: insights into mind and brain (pp. 561-572). Elsevier.

Underwood, G., Templeman, E., Lamming, L., & Foulsham, T. (2008). Is attention necessary for object identification? evidence from eye movements during the inspection of real-world scenes. Consciousness and Cognition, 17, 159-170.

Vo, M.-H., & Henderson, J. M. (2009). Does gravity matter? effects of semantic and syntactic inconsistencies on the allocation of attention during scene perception. Journal of Vision, 9(24), 1-15.

Vo, M.-H., & Henderson, J. M. (2011). Object scene inconsistencies do not capture gaze: evidence from the flash-preview moving-window paradigm. Attention Perception & Psychophysics, 73, 1742-1753.

Walther, D., & Koch, D. (2006). Modeling attention to salient proto-objects. Neural Networks, 19, 1395-1407.

Zelinsky, G., & Murphy, G. (2000). Synchronizing visual and language processing. Psychological Science, 11(2), 125-131.

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