Anticipation in Real-World Scenes: The Role of Visual Context and Visual Memory

Moreno I. Coco
Faculdade de Psicologia, Universidade de Lisboa
Alameda da Universidade, Lisboa 1649-013, Portugal
micoco@fp.ul.pt

Frank Keller
School of Informatics, University of Edinburgh
10 Crichton Street, Edinburgh EH8 9AB, UK
keller@inf.ed.ac.uk

George L. Malcolm
Department of Psychology, The George Washington University
2125 G Street NW, Suite 304, Washington, DC 20015, USA
glmalcolm@email.gwu.edu

Abstract

The human sentence processor is able to make rapid predictions about upcoming linguistic input. For example, upon hearing the verb *eat*, anticipatory eye-movements are launched towards edible objects in a visual scene (Altmann & Kamide, 1999). However, the cognitive mechanisms that underlie anticipation remains to be elucidated in ecologically valid contexts. Previous research has, in fact, mainly used clip-art scenes and object arrays, raising the possibility that anticipatory eye-movements are limited to displays containing a small number of objects in a visually impoverished context. In Experiment 1, we confirm that anticipation effects occur in real-world scenes and investigate the mechanisms that underlie such anticipation. In particular, we demonstrate that real-world scenes provide contextual information that anticipation can draw on: when the target object is not present in the scene, participants infer and fixate regions that are contextually appropriate (e.g., a table upon hearing *eat*). Experiment 2 investigates whether such contextual inference requires the co-presence of the scene, or whether memory representations can be utilized instead. The same real-world scenes as in Experiment 1 are presented to participants, but the scene disappears before the sentence is heard. We find that anticipation occurs even when the screen is blank, including when contextual inference is required. We conclude that anticipatory language processing is able to draw upon global scene representations (such as scene type) to make contextual inferences. These findings are compatible with theories assuming contextual guidance, but posit a challenge for theories assuming object-based visual indices.

Keywords: anticipation in language processing; contextual guidance; visual world; blank screen paradigm; eye-tracking.

Cognitive Science 40: 8, 1995–2024, 2016.

Introduction

The human sentence processor is able to comprehend spoken and written language with remarkable speed and accuracy: average speech rate is about 150 words per minute, while average reading speed is in the region of 200 words per minute. In order to achieve such efficiency, the sentence processor performs language understanding incrementally: as soon as a new word appears, current linguistic representations are updated to integrate the new input. Furthermore, there is increasing evidence that the sentence processor is also *predictive*, i.e., it anticipates upcoming linguistic material before it is heard or read. Prediction is presumably a key reason for the processor's efficiency, allowing it to keep up with spoken and written input as it rapidly unfolds.

A substantial body of experimental evidence in support of prediction exists. Some of this work uses the Visual World Paradigm (VWP, Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995), in which participants' eye-movements are recorded while they view a scene and listen to a spoken sentence at the same time. Fixations on the scene indicate which interpretation a listener pursues at the current point in the spoken sentence. Prediction manifests itself in the form of *anticipatory eye-movements*, i.e., fixations on regions of the scene that relate to upcoming linguistic material. Anticipatory eye-movements were first demonstrated by Altmann and Kamide (1999), who showed that listeners can predict the complement of a verb based on its selectional restrictions.[1] Crucially, anticipatory mechanisms were also observed when the sentence was listened to after the supporting visual context was removed (the blank-screen paradigm, Altmann, 2004). Participants of Altmann's (2004) study heard sentences such as (1) while viewing images such as Figure 1(a).

(1) a. The man will *eat* the cake.
b. The man will *move* the cake.

Altmann (2004) observed an increased number of saccadic eye-movements to the quadrant of the scene containing the cake during the word *eat* compared with the control condition, i.e., during the word *move* in sentence (1-b): only the cake is edible, but both objects in the scene are movable. This indicates that (1) selectional preference information provided by the verb is used as soon as it is available (i.e., incremental processing takes place), (2) this information also triggers the prediction of an upcoming argument of the verb, viz., a noun phrase that refers to an edible objects; and crucially, (3) anticipatory eye-movements can occur in the absence of a supporting visual context.

The visual world literature has shown that a wide range of linguistic information trigger anticipatory eye-movements, including case marking (Kamide, Scheepers, & Altmann, 2003), thematic role information (Kamide, Altmann, & Haywood, 2003; Kukona, Fang, Aicher, Chen, & Magnuson, 2011), tense information (Altmann & Kamide, 2007), verb subcategorization (e.g., a transitive verb triggers the anticipation of a direct object, whereas an intransitive verb does not, Arai & Keller,

[1] An alternative explanation for Altmann and Kamide's (1999) findings is through semantic association between the verb (e.g., *eat*) and the target noun (e.g., *cake*). However, this explanation can be ruled out based on subsequent work, which replicated Altmann and Kamide's (1999) anticipation effect for German, showing that looks to an upcoming noun are driven by the case marking (nominative or accusative) that noun is expected to have (Kamide, Scheepers, & Altmann, 2003).

All authors have equally contributed to the study.

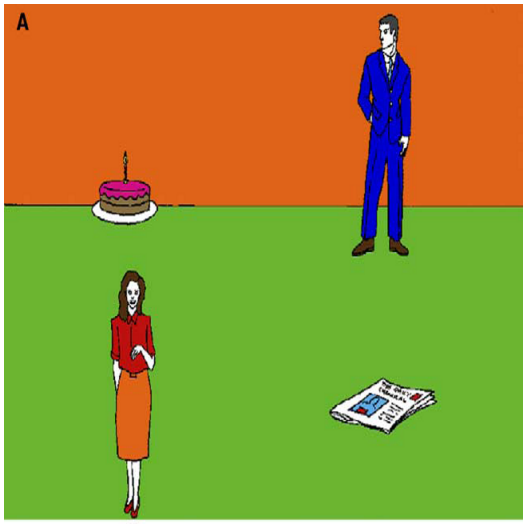
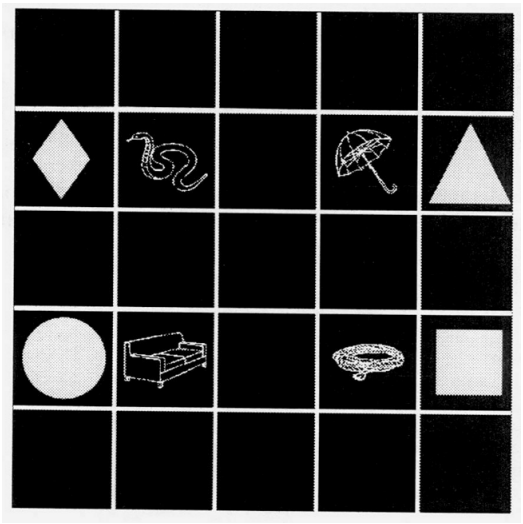
(a)  (b) 

Figure 1. Example images used as stimuli in visual world paradigm experiments. (a) clip-art scene taken from Altmann (2004), (b) object array taken from Dahan and Tanenhaus (2005)

2013), as well as non-linguistic information such as world knowledge (Knoeferle & Crocker, 2006) or visual saliency (Coco & Keller, 2015b).

Linguistic anticipation, however, is not limited to speech that is presented together with a visual context. Anticipation has also been found in reading, e.g., Staub and Clifton (2006) showed that following the word *either* readers predict the conjunction *or* and the complement that follows it; processing was facilitated compared to structures that include *or* without *either*. A range of linguistic anticipation effects have also been demonstrated in the ERP literature (for an overview, see Federmeier, 2007); again, these studies typically use spoken or written stimuli, but provide no visual context. This includes the finding that the human language processor is able to predict semantic features of an upcoming noun (Federmeier & Kutas, 1999), evidence for the anticipation of morpho-syntactic features such as gender (van Berkum, Brown, Zwitserlood, Kooijman, & Hagoort, 2005), and the demonstration that even phonological properties of an upcoming word (whether it starts with consonant or not) are predicted based on sentential context (DeLong, Urbach, & Kutas, 2005). The ERP literature has also been able to show that linguistic prediction is distinct from effects of plausibility (Federmeier, Wlotko, De Ochoa-Dewald, & Kutas, 2007) and association (Lau, Holcomb, & Kuperberg, 2013).

While there is clear evidence that language processing includes the anticipation of upcoming material, the mechanisms and representations that underlie anticipation are not well understood. In particular, although mechanisms of syntactic anticipation can be accurately captured in recent modeling work on syntactic parsing (Demberg, Keller, & Koller, 2013), it is yet to be understood how the sentence processor uses contextual information (be it visual or linguistic) to drive anticipation. In this study, we will focus on the role that visual context plays in enabling linguistic anticipation; in particular, we will investigate if visual context can be utilized by the language processor for prediction when no object information is available.

In most visual world experiments (but see Andersson, Ferreira, & Henderson, 2011; Staub, Abbott, & Bogartz, 2012), the visual context is an object array containing a small number of clearly

identifiable objects, usually arranged in a grid (see Figure 1(b) for a typical example). Alternatively, previous work has used clip-art scenes as stimuli (see Figure 1(a) for a typical example), in which the objects are arranged in a more meaningful way. However, object placement is often unnatural (both the CAKE and the NEWSPAPER are on the ground in this example), and] there is minimal] perspective (the WOMAN is in foreground, while the MAN is in the background, but their sizes are similar). Furthermore, while clip-art image content is a product of local elements, scene processing relies more on global properties (e.g., Greene & Oliva, 2009).

Eye-movements and speech are tightly time-locked (e.g., Griffin & Bock, 2000; Coco & Keller, 2012, 2015a), indicating that the cognitive system is able to efficiently link visual information to linguistic input. This is a straightforward task in impoverished visual contexts such as the clip-art scene in Figure 1(a), where there are only four objects, making it easy to identify and fixate the edible object CAKE once the verb *eat* has been processed. Objects in clip-art scenes or object arrays are not independently motivated by scene context, they are only there to provide referential anchors for the speech the participant hears. It is therefore plausible to assume that listeners develop anticipatory strategies in such a setting. In other words, it could be that the anticipation effects reported in the visual world literature are not an instance of general context-based linguistic prediction (see (Federmeier, 2007) for an overview), but are merely an artifact of the use of simple clip-art scenes as stimuli, which enables fast visual search. However, evidence against this hypothesis has been provided by Staub et al. (2012), who successfully replicated the anticipation effects of Altmann and Kamide (1999) using photographic stimuli.[2]

Based on Staub et al.'s (2012) results, we can conclude that anticipation effects can be observed in real-world scenes; they are not just artifacts of the use of clip-art scenes and object arrays that provide a small set of decontextualized referents. However, Staub et al. focused on replicating existing anticipation effects, but did not elucidate the *mechanisms* that underlie anticipation in real-world scenes. Real-world scenes differ in crucial ways from clip-art scenes or object arrays. Firstly, clip-art typically contain only a small number of objects (see the examples in Figure 2). Real-world scenes are typically more cluttered, i.e., contain a large number of objects, including groups of objects, small items, and amorphous background objects such as sky or water. The visual cognition literature shows that clutter influences search behavior, with clutter measures such as feature congestion (Rosenholtz, Li, & Nakano, 2007) correlating with search accuracy and fixation behavior (Henderson, Chanceaux, & Smith, 2009). It is therefore reasonable to assume that clutter has an effect on language-mediated eye-movements as they occur in the visual world paradigm [3], as recently shown by Coco and Keller (2015a).

The second, and perhaps most important, difference between clip-art images and real-world scenes is that the latter depict semantically coherent collections of objects, providing the viewer with contextual information which he/she can use to pro-actively allocate visual attention (see Figure 2 for an example). Contextual information restricts where objects typically occur (a table cannot float in the sky, a sofa is likely to be found indoors), as well as how objects relate to each other (monitors and keyboards tend to occur together, and in certain spatial configurations). The impor-

[2] Apfelbaum, Blumstein, and McMurray (2011) and Chen and Mirman (in press) show that the semantic relatedness of the objects depicted in a visual world display modulates lexical access, interacting with factors such as phonological neighborhood size. This result is potentially relevant in the present context, as photographic scenes, unlike clip-art arrays, are likely to contain semantically related objects, by virtue of being semantically coherent.

[3] We also tested this hypothesis on the scenes used for the experiments reported in the present paper, by correlating the number of objects in each scene with the empirical logit of fixations on the target region. This correlation was non-significant for both experiments, as detailed in the Supplementary Material to the present article.

tance of contextual information for visual tasks such as object search has been well documented experimentally (Castelhano & Henderson, 2007; Eckstein, Drescher, & Shimozaki, 2006; Ehinger, Hidalgo-Sotelo, Torralba, & Oliva, 2009; Malcolm & Henderson, 2010; Neider & Zelinsky, 2006).[4] It is therefore natural that anticipation in real-world scenes utilizes contextual information, and the aim the present study is to elucidate the mechanisms that underlie context-based anticipation. We therefore go beyond Staub et al., whos replicated the results of Altmann and Kamide (1999) in a more ecologically valid setting.

Visual context enables the listener to identify (and fixate) those regions in a complex scene in which an object predicted by the linguistic input can occur. For example in a scene such as the one in Figure 2, the verb *ate* triggers the prediction of an edible object as the argument of the verb. This prediction could then result in contextually appropriate regions being fixated: in our example, the table is the region on which edible objects are most likely to be located. This process is akin to what has been observed in the visual search literature: *contextual guidance* means that fixations occur preferentially in those scene regions where the search target object typically occurs, regardless of whether the target is actually present, or not, in the scene (Torralba, Oliva, Castelhano, & Henderson, 2006).

We can apply the Contextual Guidance Model as formulated by Torralba et al. (2006) in the context of visual attention to make the following prediction for language processing in real-world scenes: anticipation triggered by the linguistic input should manifest itself as looks to contextually relevant regions, independently of whether the target object is actually present in these regions. Even in the absence of a target, scene context can be used to determine where targets are likely to occur, based on the linguistic input heard so far. For example, we expect to observe anticipatory eye-movements to the TABLE at the verb *ate* both in Figure 2a (where the target object SANDWICH is present) and in Figure 2b (where the target object is absent). Such anticipation effects are not expected when the verb is the word *removed*, as in this case, the linguistic input does not allow the listener to predict a specific class of target objects and their contextually associated regions. Experiment 1 tests this prediction.

Once we have established that linguistic anticipation in real-world scenes utilizes contextual guidance, we need to determine how the relevant information is represented by the cognitive system. Previous work has used the blank-screen paradigm (Richardson & Spivey, 2000; Spivey & Geng, 2001; Hoover & Richardson, 2008) to investigate the visual representations that underlie language-mediated eye-movements. In this paradigm, participants first view a visual scene, which after a short interval is replaced by a blank screen; only then a speech stimulus is presented. The results show that in response to the speech they hear, participants launch eye-movements to positions on the screen at which relevant objects were located. This happens even though the screen is now blank, and no visual information is to be gained from such eye-movements. This finding generalizes to anticipatory eye-movements: Altmann (2004) demonstrates verb-based anticipation using the blank-screen paradigm, replicating the original Altmann and Kamide (1999) results (both studies used clip-art scenes).

The fact that language-mediated eye-movements happen even on a blank screen has been used to argue for a particular kind of representation of visual scenes. According to Richardson and Spivey (2000), Altmann (2004), and Richardson, Altmann, Spivey, and Hoover (2009), the findings from

[4] Again, this is not a limitation in principle: clip-art scenes could be set up to provide appropriate context objects as well as information about scene type (indoor vs. outdoor, rural vs. urban, etc.). We are not aware of published visual world experiments that use clip-art stimuli in this way.

the blank screen paradigm provide evidence that the visual system uses *visual indices*. The objects in a scene are accessed (e.g., fixated) through location-based indices, rather than through descriptions that identify them. Only a small number of visual indices are active at any given time (based on results from multiple object tracking studies, see Cavanagh & Alvarez, 2005, for an overview). Visual indices are postulated by a number of theories, including FINST Theory (Pylyshyn, 1989)), Object File Theory (Kahneman, Treisman, & Gibbs, 1992), and the Deictic Code Model (Ballard, Hayhoe, Pook, & Rao, 1997). The literature on the blank screen paradigm does not explicitly commit to one of these theories, but in this paper, we will follow the assumptions of Object File Theory, in which visual indices explicitly point to the properties of objects (unlike in FINST Theory, where indices point to visual feature available prior to object recognition). The key assumptions of Object File theory are summarized by Kahneman et al. (1992, p. 178) as follows:

> The object-centered approach that will be developed here emphasizes the distinction between identifying and seeing. We adopt the common notion that the visual field is parsed into perceptual objects and a relatively undifferentiated perceptual ground. We then assume that the main end product of perceptual processing of a stationary scene is a set of object files, each containing information about a particular object in the scene. Each object file is addressed by its location at a particular time, not by any feature or identifying label. It collects the sensory information that has so far been received about the object at that location. This information can be matched to stored descriptions to identify or classify the object, but it need not be. [...] A file is kept open so long as its object is in view, and may be discarded shortly thereafter.

In this view, scene information is accumulated at the object level, and indexed by object locations. Global scene information is not represented explicitly, which entails that contextual information (e.g., the typical location of edible objects in an indoor scene) should not be readily available in Object File Theory.

In contrast, the Contextual Guidance Model of Torralba et al. (2006) does not presuppose object representations; rather, fixation locations are computed as the product of saliency (i.e., pixel-based visual prominence) and a probability distribution over object locations conditioned on scene type, a global representation of the kind of scene being viewed (indoor vs. outdoor, etc.). Importantly, neither saliency nor scene type are object-based concepts. More formally, the key quantity in the Contextual Guidance Model is $p(X|O,G)$, i.e., the probability of fixating location X conditioned on the search target O and the scene type G (see Torralba et al. (2006), for details).

We therefore can derive two opposing predictions for language-mediated anticipatory eye-movements on a blank screen after a preview phase involving naturalistic scenes. Object File Theory predicts anticipatory eye-movements when the target object was present during preview, but not when it was absent during preview and would have to be inferred based on scene context. Such inference is not possible, as object files are not created for objects that were not observed during preview. And as the screen is blank during speech (i.e., when anticipation is computed), the cognitive system is not able to visually access contextually relevant regions from the scene; it has to rely on object files in memory.

The Contextual Guidance Model, however, makes the opposite prediction. It assumes that contextually appropriate regions can be computed based on saliency and scene type alone, which means that possible target locations can be predicted without the need for object representations, which is possible even when the screen is blank. All we have to assume is that scene type is stored

(a)  (b) 

Figure 2. Example for the experimental stimuli used in Experiments 1 and 2: (a) object present condition, (b) object absent condition, for the linguistic stimulus *the man ate/removed the sandwich*. Each scene was fully annotated with polygons using LabelMe toolbox (Russell, Torralba, Murphy, & Freeman, 2008). Highlighted in red on the scene the region of interest TABLE, contextually related to object SANDWICH. As the object SANDWICH is included within the object TABLE, we created the contextually related region on which fixation data is analyzed by merging two polygons.

in memory and retrieved during language processing; this is sufficient to compute $p(X|O,G)$, i.e., to determine the potential locations of objects anticipated by the speech. The Contextual Guidance Model, unlike Object File Theory, therefore predicts that language-mediated anticipatory eye-movements on a blank screen occur even when the relevant objects are absent during preview. We will test this prediction in Experiment 2.

Experiment 1

The aim of this experiment is to test whether anticipatory eye-movements can utilize contextual guidance. The Contextual Guidance Model predicts that viewers are able to determine likely positions of objects based on visual context. We therefore expect participants to launch language-mediated anticipatory eye-movements even in the absence of relevant target objects. In this case, viewers should fixate contextually appropriate regions (e.g., the TABLE in response to the verb *eat*). Note that contextual guidance is a feature of real-world scenes; both object arrays and clip-art scenes provide only an impoverished context that is unlikely to be sufficient for guidance to occur (see Figure 1). This aspect of real-world scenes was not tested in the previous literature (Staub et al., 2012, only used stimuli in which the target object was depicted).

The design of the present experiment closely follows Altmann and Kamide (1999), except that we use photographic scenes instead of clip-art scenes. Like Altmann and Kamide, we manipulate the verb used in the linguistic stimulus that is presented concurrently with the scene: it can be either specific (i.e., allow prediction), or ambiguous (i.e., not allow prediction). In addition, we manipulate whether the target object (e.g., the object SANDWICH in our running example) is present in the scene or not, allowing us to test whether contextual guidance occurs.

Method

Participants. Twenty-four students at the University of Edinburgh (15 females; ages 19–29, mean = 21.75) gave informed consent before taking part in this study. They were each paid £4 for participation.

Materials. The experiment used a full factorial design that crossed the factors *Verb* (Specific, Ambiguous) and *Object* (Present, Absent). The image materials consisted of 24 photo-realistic scenes, each in two versions corresponding to the two Object conditions. In the Object Present condition, we pasted a target object (e.g., SANDWICH) and a distractor object (e.g., OAR) into the scene using Photoshop. The distractor was introduced to increase the uncertainty about the target, thus preventing participants from developing looking strategies (e.g., always looking at the photoshopped object). In the Object Absent condition, neither the target and nor the distractor object were shown. Figure 2 gives an example of a scene.

In some studies using the visual world paradigm, the depicted agents seem to be looking at a particular object in the scene (see Figure 1(a)). This may induce unwanted anticipation, given that viewers are known to gaze-follow (e.g., Friesen & Kingstone, 1998; Staudte & Crocker, 2011). To avoid this confound in the present study, the depicted agents in our scenes were positioned such that they looked directly at the viewer.

Each scene was concurrently presented with a spoken transitive sentence in which we manipulated the predictivity of the Verb (Specific vs. Ambiguous). In the Specific Verb condition, the verb was highly predictive of a certain type of object (e.g., *ate* is predictive of edible objects such as SANDWICH). In the Ambiguous Verb condition, the verb was ambiguous with respect to the type of its argument (e.g., *removed* could apply a wide range of objects, not just to edible ones). A pair of example sentences is given in (2) for the scene in Figure 2.

(2) a. The man ate the sandwich.
b. The man removed the sandwich.

The full list of experimental sentences can be found in the appendix. Sentences were read by a female speaker of British English at normal speech rate without intonational breaks and recorded in a sound-attenuated studio with a Neumann km150 microphone using the Pro Tools software suite.

Predictivity Norming. Our experimental design crucially relies on the contrast between specific and ambiguous verbs. To test the validity of this manipulation, we conducted a sentence completion study in which participants were presented with the experimental sentences, but with the target noun removed, as in the following example:

(3) a. The man ate the . . .
b. The man removed the . . .

Participants saw the sentence fragments together with the corresponding images (see Figure 2 for the images corresponding to (3)), either in the Object Present condition or in the Object Absent condition. In response, they needed to type a noun that completes the sentence and refers to an object in the image. The study was conducted with 45 participants on Amazon Mechanical Turk, who were each paid $0.50. Each participant saw a list of 24 stimuli, each consisting of a sentence/image pair in one of the four conditions (Specific vs. Ambiguous Verb, Object Present vs. Object Absent). The lists were balanced using a Latin square design and randomized individually for each participant.

Target noun			
Object	Verb	Mean	SD
Present	Ambiguous	0.23	0.28
	Specific	0.64	0.36
Absent	Ambiguous	0.06	0.16
	Specific	0.43	0.36

Context noun			
Object	Verb	Mean	SD
Present	Ambiguous	0.05	0.11
	Specific	0.05	0.12
Absent	Ambiguous	0.16	0.20
	Specific	0.11	0.17

Table 1
Predictivity norming: cloze probabilities for the target noun (left) and the context noun (right)

Target noun			
Factor	Coefficient	SE	*t*
(Intercept)	0.34	0.047	7.27
Object	0.19	0.046	4.01
Verb	0.39	0.046	8.53
Object:Verb	0.04	0.093	0.39

Context noun			
Factor	Coefficient	SE	*t*
(Intercept)	0.09	0.022	4.23
Object	-0.09	0.026	-3.29
Verb	-0.03	0.026	-1.05
Object:Verb	0.05	0.052	0.94

Table 2
Predictivity norming: mixed effects model of the cloze probabilities for the target noun (left) and the context noun (right). Note that $t > 2$ indicates statistical significance.

The sentence completions obtained in this way were normalized by converting them to lowercase, removing spaces and punctuation, fixing obvious spelling errors, and turning plural nouns into singular nouns. No other post-processing was performed, therefore near-synonyms (e.g., yacht and boat) other variants (e.g., burger and hamburger) were counted as distinct responses.

In order to establish how predictive of the target noun (*sandwich* in our example) the sentence fragments were, we computed the cloze probability for each sentence, i.e., the number of completions that were identical to the target noun, divided by the total number of completions. The mean cloze probability per condition is given in Table 1 (left). We find that both for the Object Present and for the Object Absent condition, the cloze probability is higher for the Specific Verb than for the Ambiguous Verb, indicating that Specific Verbs are more predictive of the target objects, as intended. This observation is confirmed by a mixed effects model, which shows a significant main effect of Verb. In addition, we find a significant main effect of Object, indicating that cloze probabilities are higher in the Object Present condition. This is unsurprising, as participants are more likely to mention objects that are visible in the image, independent of which verb is present. There is no significant interaction of Object and Verb. The details of the mixed effects model are given in Table 2 (left).

Note that each of our image stimuli contains a context region in which the target object occurs. This region can be an object (e.g., in Figure 2, the table is the context object on which the target object sandwich occurs), or it can be a more amorphous region (e.g., grass or sky). Unlike the target object, the context region is always visible, even in the Object Absent condition, and therefore can be used as region of interest (ROI) for the analysis of the eye-tracking data generated by our

visual world experiment. However, the presence of the context region creates a potential confound: it could be that participants are predicting the context region rather than target object based on the verb they hear. This could happen in particular in the Object Absent case, where the context region, but not the target object, is visually available.

To test this hypothesis, we used our completion norming data to compute the cloze probability for the context noun, i.e., the noun referring to the context region (see appendix for the list of context nouns). In our previous example, this would be the cloze probability for the noun *table*. We report the mean cloze probability per condition in Table 1 (right); the result of a mixed effects model analysis is given in Table 2 (right). We observe low cloze probabilities for the context noun across all conditions, and crucially, the mixed effects model fails to show a significant effect of Verb, which indicates that participants do not use the verb to predict the context object. We also find a significant effect of Object: participants are more likely to produce the context noun in the Object Absent condition. This is expected: the target object is absent in this condition, and therefore is mentioned less frequently, so participants instead produce more context nouns. There is also no interaction of Verb and Object. Overall, the results of this norming study indicate that the predictability of the target object varies with the verb, but the predictability of the context region does not vary with the verb. This is the manipulation we intended for our visual world study.

Plausibility Norming. It is possible, however, that predictivity is confounded with plausibility (Federmeier et al., 2007): if a sentence involving a Specific Verb as in (2-a) is more plausible than a sentence with an Ambiguous Verb as in (2-b), then this could lead to more target noun completions (higher cloze probability), independent of how predictive the verb is. We therefore conducted a rating study in which participants judged the plausibility of our experimental items.

The materials for this study were the 24 experimental sentences, both in the Specific Verb condition and the Ambiguous Verb condition (see (2) for an example). In addition, we included versions of the experimental sentences in which the target noun was replaced by the context noun, and a preposition was added if required. For example, the context noun versions of (2) are:

(4) a. The man ate at the table.
b. The man removed the table.

This resulted in a total of 48 stimuli (see appendix for the full list). The experiment also included 24 fillers, which were designed to cover the whole range of plausibility (eight were highly plausible, eight of medium plausibility, and eight implausible).

Participants read the experimental sentences (without seeing the images) and provided a judgment on a Likert scale ranging from 1 (least plausible) to 5 (most plausible). The study was conducted with 48 participants on Amazon Mechanical Turk, who were paid $0.20. Each participant saw a list of 48 stimuli, which included each of the 24 sentence types in one of the four conditions, and all 24 fillers. The lists were balanced using a Latin square design and randomized individually for each participant.

Table 3 shows the mean plausibility ratings across the four experimental conditions. We find high ratings across the board, but also observe that sentences with the Target Noun are judged as more plausible than sentences with the Context Noun. Analyzing the data with a mixed effects model (see Table 4 for details) confirms this by showing a significant main effect of Noun. The mixed effects model fails to show a significant main effect of Verb, which confirms that our sentences work as intended: while there is a difference in predictivity between Specific and Ambiguous Verbs, the two verb types do not differ in plausibility, thus dis-confounding predictivity and plausibility in

Noun	Verb	Mean	SD
Target	Ambiguous	4.24	1.11
	Specific	4.63	0.89
Context	Ambiguous	3.89	1.37
	Specific	3.56	1.52

Table 3
Plausibility norming: mean judgments for the experimental stimuli and their context-noun counterparts.

	Target noun		
Factor	Coefficient	SE	*t*
(Intercept)	4.09	0.11	36.88
Noun	0.70	0.18	4.01
Verb	0.02	0.17	0.13
Noun:Verb	0.74	0.28	2.63

Table 4
Plausibility norming: mixed effects model of the judgments for the experimental stimuli and their context-noun counterparts. Note that $t > 2$ indicates statistical significance.

our experimental stimuli.

However, Table 4 shows a significant interaction of Noun and Verb. This interaction could indicate a problem for our stimuli, if it turned out that Specific Verbs are more plausible than Ambiguous Verbs in the Target Noun condition (or vice versa in the Context Noun condition). We tested this using confidence intervals, as proposed by Masson and Loftus (2003), and applied to mixed effects models by Levy and Keller (2013). The model predicted value for the Target/Ambiguous condition was 4.25, with a model confidence interval (CI) of 0.50, while the predicted value for the Target/Specific condition was 4.64, with a CI of 0.32. The difference between predicted values indicates that the CIs overlap ($0.39 < 0.41$), which means that the two conditions are not reliably different. The predicted value for the Context/Ambiguous condition was 3.91 (CI = 0.59), while for the Context/Specific condition, it was 3.57 (CI = 0.71). Again, the difference between the predicted values indicates that the CIs overlap ($0.35 < 0.65$), hence the two conditions are not reliably different.[5]

Procedure. After having reported the studies used to norm our materials, we now turn to the main experiment, which used the visual world paradigm. An EyeLink II head-mounted eye-tracker was used to monitor participants eye-movements with a sampling rate of 500 Hz. Images were presented on a 21" multiscan monitor at a resolution of 800 $\times$ 600 pixels. Only the right eye

[5]The CI-based analysis also showed that the Target/Specific and the Context/Ambiguous conditions were reliably different (difference in predicted values 0.72, difference in CIs 0.46), as well as the Target/Ambiguous and the Context/Specific conditions (difference in predicted values 0.68, difference in CIs 0.61). These differences are not of theoretical interest, but they explain why there is a significant Noun:Verb interaction in the mixed effects model.

was tracked.

Participants were free to move their heads, though were asked to try to remain as still as possible. Participants sat approximately 60 cm away from the monitor. They were informed about the experimental procedure using written instructions, which told them that they would be looking at scenes and hearing sentences at the same time. They were asked to pay attention to both the scene and the sentence, as there would be questions after some of the trials.

The experiment started with a standard nine-point calibration of the eye-tracker, and each trial began with a drift correction in order to confirm tracking accuracy. Then a scene would appear on the screen for 1500 ms. We used 1500 ms here, rather than the 1000 ms of Altmann and Kamide (1999), as our scenes were more complicated (photorealistic rather than clip-art), and the agents were directly looking at the participant, which may capture participants' attention at the onset of a trial. After the preview, the sentence relating to the image was played; at the end of it, participants were given an extra 1500 ms before scene offset.

Questions followed a random selection of 25% of the trials; half of the questions were about the visual specifics in the scene (e.g., *was the man near a table?*), the other half about the sentence (e.g., *did the man move the sandwich?*). The questions were there to ensure that participants were attending to both the scene and sentence. All questions were yes/no questions and were answered with a button press. Questions were answered correctly 97.2% of the time. A logistic mixed effect model with Verb and Object as fixed effects, Participant and Item as random slopes and intercept, was used to test whether question accuracy varied with the experimental factors of interest. We found neither significant main effects nor an interaction (all $p > 0.3$).

The materials were distributed across four lists according to a Latin Square design, such that each list contained every scene in one of the four experimental conditions. Each participant saw one list, together with 96 fillers (the same for every participant). Half of the filler scenes contained photoshopped objects; each filler scene was paired with a short descriptive sentence. Each participant saw the experimental items and fillers in an individually generated randomized order. The experiment lasted approximately 30 minutes in total.

Analysis. In order to investigate verbally driven anticipatory effects, we analyze eye-movement data from 100 ms after the onset of the verb (e.g., the word *ate*), to account for the oculo-motor delay, i.e., the time needed to launch a saccade in response to the auditory input, until the beginning of the post-verbal noun (the word *sandwich* in our example). The mean duration of the verb was 476.14 ms (SD = 144.43 ms), while the post-verbal noun started at 682.89 ms on average (SD = 158.15 ms) after the verb onset. For the purposes of our analysis, a fixation is counted from the onset of the saccade leading into it.

As the onset of the verb varies across sentences, we align fixations on an item-by-item basis, i.e., we use the timestamps for the verb onset of each sentence to decide which fixations to include. Moreover, fixations launched after the end of the ROI (i.e., the onset of the post-verbal noun for this item) are excluded from the analysis (5% of all fixations on the ROI). This step ensures that fixations that are potentially contaminated by the next word (the post-verbal noun, i.e., the head of the direct object) do not contribute to the analysis. The reason for this is that fixations that happen once the direct object has been heard can no longer be regarded as predictive. Note that in our plots, we do not mark the end of the ROI, as it varies from item to item. In the Supplementary Material, we also show time-course plots of fixation across the entire sentence, and test whether fixations differ at the onset of the linguistic region of interest (a 400 ms window centered at the onset of the verb). This additional analysis serves to demonstrate that there are no differences in fixation across

experimental conditions arising prior to the processing of the verb (see Supplementary Material for details).

For the purposes of the analysis, we consider fixations on the contextually relevant region (TABLE in our example), which always includes the target object directly associated with the linguistic information (SANDWICH in our example). We are interested in the anticipation of contextually relevant regions, rather exclusively on a specific target object. Specifically, in the Object Absent condition, the actual position of the target can only be guessed by the participants, while the contextually relevant region is always depicted in our scenes.

Two annotators (blind to the purpose of the experiment) independently marked up the contextually relevant target regions. They were given the images in the Object Absent condition (e.g., Figure 2b) together with the target object (e.g., SANDWICH) and were instructed to draw a polygon around the region in the image in which the target is likely to occur. In our analysis of the eyetracking data, we used the intersection of the two polygons drawn by the annotators for each image.

For both visualization and data analysis, we use the empirical logit[6] of fixation counts, which is conceptually a log-odds ratio of the probability of fixation on the target region and the probability of fixating not on the target region (Barr, 2008). Specifically, the empirical logit is defined as $\text{emplog}(y) = \log \frac{y+0.5}{N-y+0.5}$, where y is the number of fixations on the target region and N is the total number of fixations on all the objects in a scene (including the background) within a 50 ms time window.[7] An aggregation in 50 ms window is enough to significantly reduce correlations between consecutive time points, while preserving temporal variability between points (i.e., a participant might be fixating two different objects within a 50 ms window); see Dale Barr's blog `http://talklab.psy.gla.ac.uk/tvw/agg/agg.html` for an unpublished analysis of aggregation. In the Supplementary Material, we also report an analysis of empirical logit of fixation when calculated over the whole linguistic region of interest (a 600 ms window). This analysis shows results corroborating with what was reported in the main paper.

We use mixed effects models (Pinheiro & Bates, 2000) to statistically analyze our fixation data. The mixed effects analysis uses the empirical logit as the dependent variable (12 time bins of 50 ms each), the associated weights, and the following centered predictors: *Verb* (Specific, 0.5; Ambiguous, -0.5), *Object* (Present, 0.5, Absent, -0.5). *Time* is represented as orthogonal polynomial of order two (Time^1, Time^2). The polynomial representation of Time, originally proposed by Mirman, Dixon, and Magnuson (2008), gives us a better way of capturing the temporal dynamics of fixations, and returns a more accurate estimate of the model fit. The main reason for using this approach is that fixations almost never distribute linearly in time; fitting a polynomial allows us to estimate non-linear changes in fixation probability across time. We use a polynomial of order two. The linear term of the polynomial has exactly the same interpretation as a linear regression of fixations over time, e.g., a positive Time^1 indicates an increase of fixation over time. The quadratic term can be used to identify sudden changes in the linear trend and characterize the 'bowing' of the fixation function. For example, a positive Time^2 coefficient indicates the curvature of the fixation function is upward, i.e., an increase followed by a decrease. Higher order terms improve the model

[6] We also modeled fixations using logistic mixed effects models, thus keeping fixation as binary (coded as 0 and 1). We observe the same trend of effects but at the cost of statistical power: information is lost because we are treating each time window as binary. To increase statistical power, we could consider fixations at their sample rate, i.e., 2 ms, but this would highly increase their temporal correlation, which is due to fixations not being independently sampled. For this reason, we decided to use empirical logit as our dependent measure.

[7] We also calculate the weights of the empirical logit as $(\frac{1}{y+.5} + \frac{1}{N-y+.5})$, and include it in the mixed effects models to account for potential variance across observations.

fit but are difficult to interpret (Mirman & Magnuson, 2009). The random effects of our analysis are Participants and Items.[8]

We fit full mixed effects models, i.e., models that include all fixed effects and their interactions, with a maximal random structure, in which random variables are included both as random intercepts (e.g., `(1 | Participants)`, using the `lme4` syntax) and as uncorrelated random slopes (e.g., `(0 + Verb | Participants)`). This approach is known to result in the lowest rate of Type 1 error (Barr, Levy, Scheepers, & Tily, 2013). Object, Verb and Time are all included as random slopes. Note, however, that we do not introduce interactions as random slopes (e.g., `(0 + Verb:Object | Participants)`), as the resulting models do not converge.

In the Results section, we visualize the observed empirical logits for the different experimental conditions as shaded bands indicating the standard error around the mean, and overlay the model fits as lines. In the results tables, we report the coefficients and standard errors of the mixed effects models, and derive p-values from the t-values (also reported) for each of the factors in the model. The t-distribution converges to a z-distribution when there are enough observations, and hence we can use a normal approximation to calculate p-values.

Results

In Figure 3, we plot the time course of fixations, represented as the empirical logit of fixation probabilities and drawn as a shaded band, from 100 ms until 700 ms, which corresponds exactly to the verb region of interest (see Analysis section for details on how the ROI was defined on an item-by-item basis). The figure also includes the estimated values generated by the mixed effects model (plotted as smooth curves).

If the target object is present, we observe a clear anticipation effect: looks in the Specific Verb condition increase more rapidly than looks in the Ambiguous Verb condition (i.e., the blue curve rises more steeply than the red curve), indicating that participants are able to use the verb information to anticipate the target object (corresponding to the post-verbal noun in the sentence). A similar trend is observed in the target absent condition; again looks in the Specific Verb condition increase more steeply than in the Ambiguous Verb condition, though the increase is perhaps less steep than in the Object present condition. The mixed effects analysis in Table 5 is consistent with this: we find a significant interaction of Verb and Time^1 with a positive coefficient, indicating that looks increase more steeply over time in the Specific Verb condition than in the Ambiguous Verb condition, indication the prediction effect. This replicates Altmann and Kamide's (2009) classic result. We also find a marginal interaction of Verb and Time^2, indicating that this increase may have a quadratic component. We also find a significant main effect of Time^2, indicating that there is a quadratic increase of looks to the target over time, independent of condition, and a significant interaction of Object and Verb, indicating there are more looks overall to Specific Verbs in the Object Present condition. Crucially, however, there is are no significant interactions involving Object and Time, showing that the presence or absence of the target object does not influence how looks to the target region develop over time. In particular, the presence or absence of the target object has no

[8] We include both random variables in our model, rather than aggregating by participants and by items in order to arrive at a single model accounting for the variance of both random factors simultaneously. Note that the solution of fitting separate mixed effects models with fixation data aggregated by participants and items is reminiscent of the issue of computing separate Anovas (by participants and by items), which makes it hard to establish the significance of a factor that is found significant by participants but not by items.

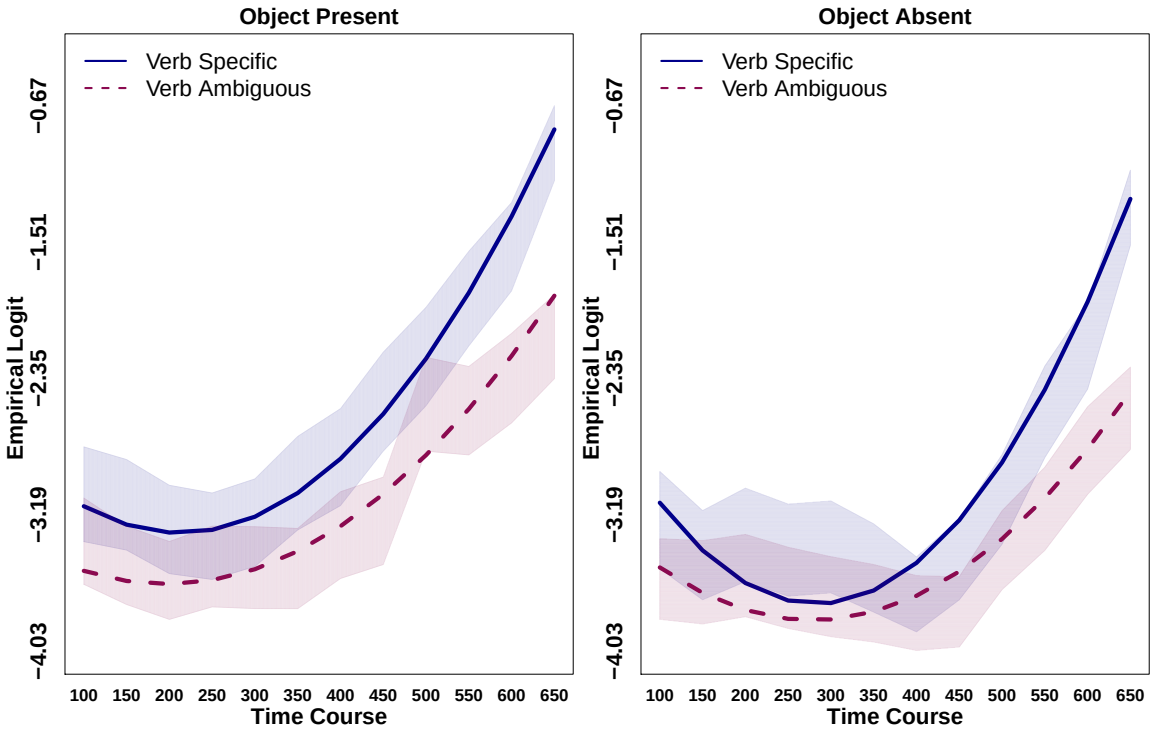


Figure 3. Experiment 1: Time course plot of the empirical logit of fixations from 100 ms to 700 ms after the onset of verb for the different experimental conditions on the target region. Left panel: Object Present condition, right panel: Object Absent condition. The shaded bands indicate the standard error around the observed mean. The lines represent the predicted values of the mixed effects model reported in Table 5. Note that the offset of the region of analysis varied by items, but fixations crossing the offset were excluded, see Analysis section for details.

bearing on anticipation; if this was the case, then we would have observed an interaction of Object, Verb, and Time.

Discussion

This experiment had two main results. Firstly, it provided evidence that linguistically driven anticipation effects can be observed in real-world scenes. In other words, anticipation is not just a consequence of restricting the visual scene to a small number of decontextualized objects, as most previous work using the visual world paradigm has done. Rather, we found that anticipatory eye-movements happen even in photographic scenes: which are often cluttered, contain a visually diverse set of objects, and do not restrict what can be talked about in a predictable fashion. More specifically, we observed that looks to the target region increase more quickly in the Specific Verb condition than in the Ambiguous Verb condition (see Figure 3), as evidenced by the significant interaction of Verb and Time[1] (see Table 5). This indicates that participants are able to predict the direct object of a verb based on the selectional restrictions of that verb, and launch anticipatory eye-movements to the region in which the direct object is likely to be depicted. This result replicates the

Table 5
Experiment 1: Results of the mixed effects model analysis. The dependent measure is empirical logit of the fixation probability on the target region; predictors are Time (100 to 700 ms, in 50 ms intervals) represented as an orthogonal polynomial of order two (Linear, Time^1; and Quadratic, Time^2), Verb (Specific, 0.5; Ambiguous, −0.5) and Object (Present, 0.5; Absent, −0.5). Random intercepts of Participants and Items and random slopes for main effects were included in the model, see Analysis section for details.

Predictor	β	*SE*	*t*	*p*
Intercept	-3.06	0.17	-17.78	<0.0001
Object	0.46	0.28	1.61	0.1
Verb	0.39	0.25	1.54	0.1
Time^1	1.97	0.38	5.15	<0.0001
Time^2	1.13	0.20	5.52	<0.0001
Object:Time^1	0.39	0.23	1.69	0.09
Object:Time^2	-0.32	0.23	-1.41	0.15
Verb:Time^1	0.71	0.23	3.10	0.001
Verb:Time^2	0.56	0.23	2.46	0.01
Object:Verb	0.21	0.13	1.54	0.1
Object:Verb:Time^1	-0.17	0.46	-0.39	0.6
Object:Verb:Time^2	-0.39	0.46	-0.85	0.3

findings of Staub et al. (2012), who demonstrated verb-specific anticipation effects using photorealistic scenes. Our results extend also earlier work by Andersson et al. (2011), which showed that eye-movements are time-locked with speech in photorealistic scenes. They observed large latencies from mentioning an object to fixating it, leading them to conclude that "it seems unlikely that effects of anticipatory eye-movements will have adequate time to occur" in photorealistic scenes (p. 214). This claim is not consistent with the results of our Experiment 1, where we indeed observe anticipation. A possible reason for this discrepancy is that Andersson et al.'s photographs were more cluttered than the ones that we used (Henderson et al., 2009). However, when we tested this hypothesis in our data, we failed to find a significant correlation between the number of objects in a scene and the empirical logit of fixations on the target object, see Supplementary Material for details.

The second finding of this experiment was that anticipatory eye-movements triggered by the verb happen both when the target object is depicted in the visual scene and when it is absent. This result provides evidence for *contextual guidance* in anticipatory eye-movements. As argued in the Introduction, real-world scenes, while being visually more complex, contain important information that clip-art scenes lack: they provide visual context (e.g., scene type, co-occurring objects) that makes it possible to compute where an object is likely to be located. Context can make it easier, rather than harder, to find objects in a scene, as a growing body of work investigating visual search in real-world scenes has shown (Torralba et al., 2006; Castelhano & Henderson, 2007; Eckstein et al., 2006; Ehinger et al., 2009; Malcolm & Henderson, 2010; Neider & Zelinsky, 2006).

Note, however, that what we demonstrated in this experiment goes beyond previous results

on contextually guided visual search. Unlike visual search experiments, the present experiment provided no explicit search target. Rather, listeners used the linguistic input to anticipate what type of object would be mentioned next (e.g., an edible object in the case of the verb *eat*) and then initiated visual search for such an object. This search was subjected to contextual guidance, as the search target was not explicitly specified at the point of hearing the verb. This also significantly differentiates our results from previous work demonstrating anticipation effects using the visual world paradigm, which almost always used an explicitly depicted target object, both with clip-art stimuli (e.g., Altmann & Kamide, 1999) and with real-world stimuli (Staub et al., 2012); see, however, Altmann and Kamide (2009) for a clip-art study where the target object is not explicitly mentioned.

Experiment 2

In Experiment 1, we found anticipation based on the selectional restrictions of the verb, both when the target object is depicted in the visual scene and when it is absent. This is in line with a key prediction of the Contextual Guidance Model: language-mediated anticipatory eye-movements can target locations based on scene context, rather than explicitly specified in the linguistic input. This finding raises the question of which visual representations drive anticipatory eye-movements. Our second experiment investigates this issue by using the *blank screen paradigm*, in which a visual scene is previewed, replaced by a blank screen, and then eye-movements are recorded on the blank screen while the linguistic stimulus is presented. Participants therefore have to rely on memory representations of the scene when making language-mediated eye-movements, as they cannot access the visual information any more when hearing the sentence. That anticipation happens in this set-up was demonstrated using clip-art scenes by Altmann (2004).

In the Introduction, we identified two contrasting hypotheses regarding the visual representations relevant for linguistic anticipation. Object File Theory (Kahneman et al., 1992) assumes that visual representations are object-based, such that each object is stored as a collection of properties pertaining to it, together with a visual index that anchors it in the visual scene. As object files are only created for objects that are present in the scene, we predict that anticipatory eye-movements on a blank screen occur only if the target object is explicitly depicted during preview; if the target object is absent during preview, then no object files are available to relate linguistic prediction to.

In contrast to this, the Contextual Guidance Model of Torralba et al. (2006) assumes that contextually appropriate regions (in visual search) are computed based on global features of the scene (scene type), without recourse to explicit object representations. This model therefore predicts that contextual guidance should occur even if the target object is absent during preview, under the assumption that scene type information is stored in memory, and is thus available to compute contextually appropriate regions even on a blank screen.

The present experiment tests the predictions of both theories by replicating Experiment 1 using the blank screen paradigm. If object file theory is correct, then we expect to see anticipatory eye-movements on the blank screen only when the target object is present during preview; if the Contextual Guidance Model is correct, then we predict anticipation on the blank screen both when the target object is present and when it is absent.

Method

Experiment 2 closely follows the design of Experiment 1, i.e., it crosses the factors *Verb* (Specific, Ambiguous) and *Object* (Present, Absent). The same 24 experimental items and 96 fillers were used (both images and speech stimuli were re-used). We recruited twenty-four new participants from the same population as in Experiment 1.

Experiment 2 differed in experimental procedure from Experiment 1 as follows: Instead of seeing the scene and hearing the sentence concurrently, participants now previewed the scene for 5000 ms, then a blank screen for 1000 ms, after which the sentence started to play (while the screen remained blank). This procedure follows closely the one used by Altmann and Kamide (2009), who also used a 5000 ms preview and a 1000 ms interval between image and speech.

Again, 25% of the trials included questions; these were answered correctly 87.5% of the time. Again, a logistic mixed effects model with Verb and Object as fixed effects, Participant and Item as random slopes and intercept, was used to whether the experimental factors Object and Verb affected question accuracy. As in Experiment 1, we found neither significant main effects nor an interaction (all $p > 0.3$).

We analyze the data gathered in this experiment using the empirical logit of fixation proportions on the target region, as in Experiment 1. The procedure for the mixed effects analysis was the same as in Experiment 1: we included Verb (Specific, Ambiguous), Object (Present, Absent) and Time (Linear, Quadratic) as fixed factors, and Participants and Items as random effects. The dependent variable was the empirical logit of fixation proportions on the target region, and it was weighted. We used a full model with maximal random structure, as outlined in the Analysis section of Experiment 1. The definition of the region of interest for analysis was also the same as in Experiment 1, incorporating all fixations in the 100–700 ms interval after the verb onset for each item, except for fixations that cross the next ROI boundary; i.e., we only consider fixation data related to the verb. The reader is referred to the Supplementary Material for corroborating results obtained with a single 600 ms window.

Results

In Figure 4, we plot the empirical logit of the fixation probabilities (shaded bands), together with the estimated values predicted by the mixed effects model (smooth curves). As in Experiment 1, we plot the interval of 100–700 ms from the verb onset (see Analysis section for a precise definition of the region of interest).

In the Object Present condition, we again find a clear anticipation effect: looks to the target region increase more quickly in the Specific Verb condition than in the Ambiguous Verb condition, consistent with the assumption that participants use verb selectional restrictions to predict the direct object of the verb before they have heard it. In the Object Absent condition, we observe virtually the same anticipation effect. This observation is consistent with the mixed effects analysis, whose results are given in Table 6. We find a significant interaction of Verb and Time^1, as well as a significant interaction of Verb and Time^2, indicating that looks to the target increase more steeply for Specific Verb, and that this increase has both a linear and a quadratic component. As in Experiment 1, we find a main effect of Time^1 and Time^2 (looks increase over time, independent of condition), and an interaction of Object and Verb, which suggests that in the Object Absent condition, there are more anticipatory looks when the Verb is Specific. This effect, however, does not change over time. Again, we fail to find any interactions involving both Object and Time, consistent with the

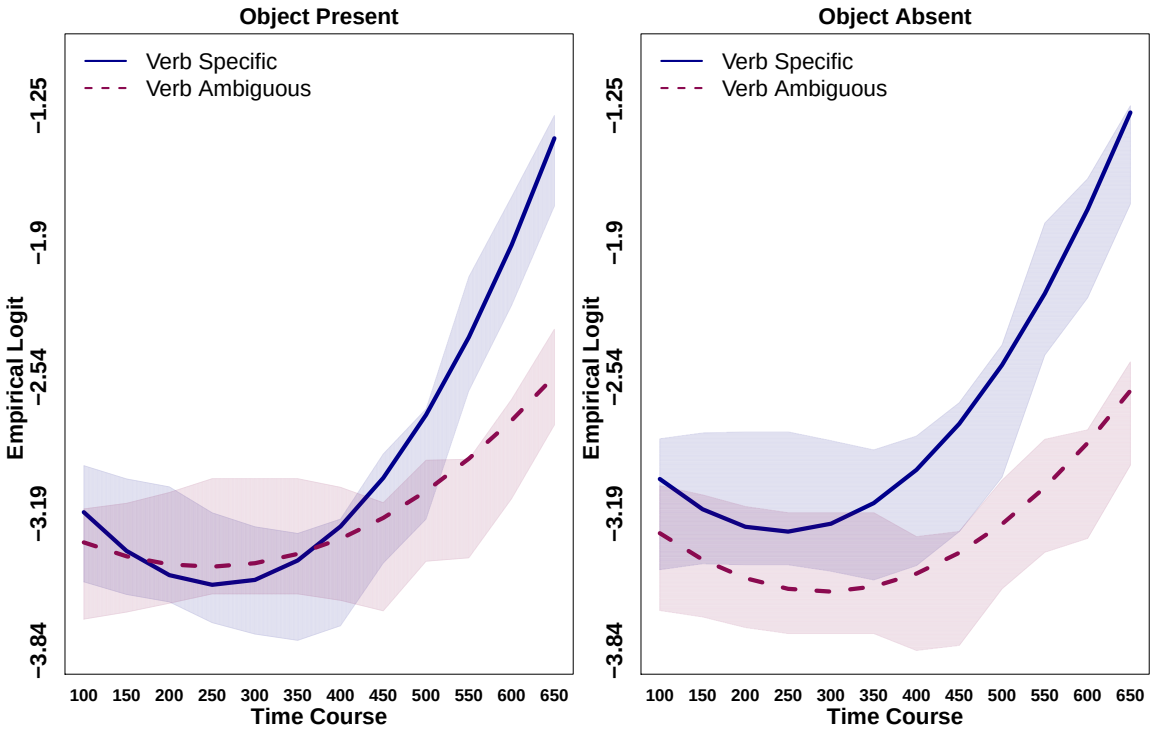


Figure 4. Experiment 2: Time course plot of the empirical logit of fixations from 100 ms to 700 ms after the onset of verb for the different experimental conditions on the target region. Left panel: Object Present condition, right panel: Object Absent condition. The shaded bands indicate the standard error around the observed mean. The lines represent the predicted values of the mixed effects model reported in Table 6. Note that the offset of the region of analysis varied by items, but fixations crossing the offset were excluded, see Analysis section for details.

observation that the presence or absence of the target object has no influence on the time course of fixations. Specifically, the three-way interactions Object:Verb:Time^1 and Object:Verb:Time^2 are not significant, which means that the presence or absence of the target object does not modulate the verb-based anticipation effect.

Discussion

The present experiment demonstrated that linguistically-driven anticipation can occur on a blank screen, even if the scene previewed by participants is a real-world scene. Our results therefore replicate and extend the findings by Altmann (2004), who demonstrated anticipation on a blank screen using clip-art scenes. On a general level, our experiment showed that the memory representations that underpin anticipation in clip-art scenes are also available in real-world scenes (see Ferreira, Apel, & Henderson, 2008, and Richardson et al., 2009, for further discussion of the representation issue).

Extending previous work, the results of Experiment 2 put theoretical constraints on the type of representations that underlie anticipatory eye-movements. We found an interaction of Verb and

Table 6

Experiment 2: Results of the mixed effects model analysis. The dependent measure is empirical logit of the fixation probability on the target region; predictors are Time (100 to 700 ms, in 50 ms intervals) represented as an orthogonal polynomial of order two (Linear, Time^1; and Quadratic, Time^2), Verb (Specific, 0.5; Ambiguous, −0.5) and Object (Present, 0.5; Absent, −0.5). Random intercepts of Participants and Items and random slopes for main effects were included in the model, see Analysis section for details.

Predictor	β	*SE*	*t*	*p*
Intercept)	-3.07	0.32	-9.56	<0.0001
Object	0.01	0.25	0.07	0.9
Verb	0.41	0.28	1.45	0.1
Time^1	1.47	0.43	3.37	0.0007
Time^2	0.90	0.2	4.45	<0.0001
Object:Time^1	-0.13	0.2	-0.64	0.5
Object:Time^2	-0.003	0.2	-0.01	0.9
Verb:Time^1	1.1	0.2	5.33	<0.0001
Verb:Time^2	0.62	0.2	3.005	0.002
Object:Verb	-0.29	0.12	-2.35	0.01
Object:Verb:Time^1	0.11	0.41	0.26	0.7
Object:Verb:Time^2	0.34	0.41	0.84	0.3

Time, but no interactions involving Object and Time, which means that verb-based anticipation happens independently of whether the target object is present during preview or not (see Figure 4). As we argued in the Introduction, this result is expected if we assume that contextual guidance is involved in anticipatory eye-movements: Torralba et al.'s (2006) Contextual Guidance Model assumes that scene type is used to compute contextually appropriate target regions for the visual search (in our case, this visual search is triggered by verb-based anticipation). According to the Contextual Guidance Model, this process is independent of whether the target object is actually present in the scene or not, as it does not rely on object-based representations and therefore should also occur on a blank screen (assuming that scene type is stored in memory).

On the other hand, Object File Theory is not able to predict the results of the present experiment: it assumes that anticipatory eye-movements are driven by object-based representations which are stored in memory and include visual indices that trigger fixations on the blank screen. Crucially, object files can only be formed for objects that are depicted. Object File Theory therefore predicts that anticipatory eye-movements on the blank screen should be limited to the condition in which the target object was present during preview, as object files are available only in this case. This is contrary to what we found in the present experiment.

General Discussion

We can summarize the main results of the two experiments presented in this paper as follows: Experiment 1 confirmed that linguistic anticipation can be observed in real-world scenes and showed that such anticipation effects occur even when the target object is not depicted in the scene. Experiment 2 found that participants make anticipatory eye-movements when viewing a blank screen, both when the target object is present during preview, and when it is not depicted.

As argued in the Introduction, our results constrain theories of language-mediated anticipation in important ways. Experiment 1 highlights the role of context in anticipation. Rather than hindering anticipation (as Andersson et al., 2011, claim), the visual complexity of real-world scenes makes anticipation possible in certain cases: when the target object is not depicted, participants inspect the location at which the target would typically occur given the scene context (a boat would occur on water, etc.). The use of real-world scenes is crucial to this finding, as virtually all prior visual world experiments have used simple clip-art scenes or object arrays which provide very little object context or scene type information.

We argued that the Contextual Guidance Model of Torralba et al. (2006) provides a natural explanation for the results of Experiment 1: according to this model, visual search proceeds by computing fixation locations based on visual saliency combined with a probability distribution modeling where a given type of object is likely to be located, conditioned on scene type. Scene type is a coarse-grained visual property that indicates whether a given scene is indoors or outdoors, cityscape or landscape, etc. Crucially, the Contextual Guidance Model does not assume that object-based representations are required for contextual guidance; fixation locations are computed based on salience (a property of the pixels in a scene) and scene type (a global property of the scene), together with a probability distribution over object locations. This distribution is pre-computed in a separate training phase; object locations only need to be assumed during training time, not when fixations are computed on a new scene that has not been viewed previously.

However, the results of Experiment 1 are also compatible with the assumption that anticipation involves object files, i.e., bundles of object properties that are linked to the objects themselves through location-based indices, as Richardson and Spivey (2000), Altmann (2004), and Richardson et al. (2009) have argued. We can assume that even when the target object is not depicted in a scene (as in the Object Absent condition of Experiment 1), its potential location can be inferred: once participants hear the verb *ate*, they know that the argument of the verb will refer to an edible object. They therefore perform a visual search for a location at which such an object is likely to be located, and fixate the TABLE region of the scene. In other words, likely target locations are not pre-computed as in the Contextual Guidance Model, but they are computed on the fly based on the visual information available in the scene.

The results of Experiment 2 are again compatible with the Contextual Guidance Model. In this model, contextual guidance does not rely on object representations, and thus should be available both when the target object is depicted during preview, and when it is absent. We can therefore explain why verb-mediated prediction effects were attested in Experiment 2 both in the Object Present and the Object Absent conditions.

Object File Theory, on the other hand, does not predict anticipation on the blank screen when the target object is absent from the scene. No object files are constructed for absent objects, hence no visual indices can be assigned to such objects. Furthermore, the location of the absent target object cannot be inferred by inspecting the scene, as the scene is no longer present when speech-

mediated eye-movements are made (unlike in Experiment 1, where the scene and the sentence were co-present). Object File Theory is therefore at odds with the finding of Experiment 2 that anticipation occurs on the blank screen even when the target object is absent.

We could try to rescue Object File Theory by assuming that anticipation is based on the context object (TABLE in our example), rather than on the target object (SANDWICH in our example). Unlike the target object, the context object is presented during preview before the screen goes blank, even in the object absent condition. Therefore, participants can build up an Object File for the context object, and access it (i.e., fixate its location) when they hear the verb. In other words, participants could predict *table* upon hearing *eat* and fixate the region where the TABLE was located, without the need for an object file for SANDWICH (which was absent during preview and thus cannot be constructed). However, this alternative explanation is inconsistent with our predictivity norming study (see Methods section of Experiment 1). In this study, we established that specific verbs such as *eat* are more predictive of the target object than ambiguous verbs such as *remove*. However, we also investigated the cloze probability of the context object, and found that specific verbs and ambiguous verbs do not differ in their predictivity of the context object. We should therefore observe no anticipation effects in the object-absent condition of the blank screen experiment, contrary to fact.

To summarize, the results presented in this paper provide evidence that anticipation in visually situated language comprehension is a general process that can occur even in naturalistic scenes; it is not limited to clip-art scenes with a handful of clearly distinct objects. Moreover, we showed that anticipation is able to make use of contextual information present in naturalistic scenes: in case a suitable target object is not depicted, its typical location can be determined based on visual context. Both these findings hold even when the visual scene is no longer depicted during language processing, i.e., if the blank screen paradigm is used. We argued that this finding is compatible with models that use non-object based representations of a scene, such as the Contextual Guidance Model, but not with models that rely on object files and visual indices to represent a scene.

Theoretical Alternatives

A potential way of overcoming the limitations of Object File Theory would be to augment it with an object-based (rather than scene-based) account of visual context. Such an account has been proposed by Hwang, Wang, and Pomplun (2011) in the form of *semantic guidance*. Hwang et al. (2011) use Latent Semantic Analysis (Landauer & Dumais, 1997) to compute the semantic similarity between objects in a scene, and show that during scene inspection, participants prefer to transition to objects that are semantically similar to the object currently fixated. Hwang et al. also find that during visual search, participants tend to fixate objects that are semantically similar to the search target. These findings could provide a mechanism for explaining the results of Experiment 2: Assume that the memory representations for the objects in a scene contain not only visual indices and basic object properties (e.g., color, material, size), but also a list of semantically related objects. Such a representation could explain context-driven linguistic anticipation, even when the target object is absent. For example, in Figure 2b, participants have a representation of TABLE which includes the fact that SANDWICHES are semantically related. Upon hearing *ate*, they predict an edible object, such as SANDWICH to be mentioned next, and fixate TABLE, which is semantically related. Information about semantic relatedness is part of the memory representation of TABLE, and therefore is available even when SANDWICH is not depicted and when the screen is blank. Evidence for such an association-based view comes from studies that find eye-movements to objects that

are semantically related to target words presented auditorily (e.g., Yee & Sedivy, 2006; Huettig, Quinlan, McDonald, & Altmann, 2006).

The integration of Object File Theory with semantic guidance provides a more natural, fully object-based account of the results of the two experiments presented in this paper. It also has the advantage of being compatible with visual world results that do not use real-world scenes (and still find anticipation effects, even when using the blank-screen paradigm); this is because semantic guidance does not rely on context in the form of a global representation of scene type, it only requires the objects in the display to be semantically related. Such relatedness is automatically present in coherent real-world scenes, but also the objects in a clip-art scene can trigger semantic guidance effects (Huettig & Altmann, 2005; Dahan & Tanenhaus, 2005). A downside of semantic guidance, on the other hand, is the fact that it assumes that everything that triggers guidance is represented as an object, such as SKY, GRASS, and other large background regions, which presumably are represented differently from regular objects.

Future research is required to adjudicate between contextual and semantic guidance as explanatory mechanisms underpinning linguistic anticipation effects in real-world scenes. As indicated above, the two models differ in the role they assign to global properties such as scene type. It would be possible, for instance, to devise an experiment in which scene type is manipulated (e.g., using indoor vs. outdoor scenes). Under contextual guidance, we would expect fixation locations to differ for different scene types if the target object is absent. (For example, in street scenes, pedestrians tend to be located near the bottom of the scene, while in indoor scenes, they tend to be located near the center.) Semantic guidance, on the other hand predicts no difference, provided that the non-target objects in the scene remain constant across scene types.

References

Altmann, G. T. M. (2004). Language-mediated eye movements in the absence of a visual world: the blank screen paradigm. *Cognition*, *93*, B79– B87.

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

Altmann, G. T. M., & Kamide, Y. (2007). The real-time mediation of visual attention by language and world knowledge: Linking anticipatory (and other) eye movements to linguistic processing. *Journal of Memory and Language*, *57*, 502–518.

Altmann, G. T. M., & Kamide, Y. (2009). Discourse-mediation of the mapping between language and the visual world: Eye movements and mental representation. *Cognition*, *111*, 55–71.

Andersson, R., Ferreira, F., & Henderson, J. M. (2011). The integration of complex speech and scenes during language comprehension. *Acta Psychologica*, *137*, 208–216.

Apfelbaum, K. S., Blumstein, S. E., & McMurray, B. (2011). Semantic priming is affected by real-time phonological competition: Evidence for continuous cascading systems. *Psychonomic Bulletin and Review*, *18*, 141–149.

Arai, M., & Keller, F. (2013). The use of verb-specific information for prediction in sentence processing. *Language and Cognitive Processes*, *28*(4), 525–560.

Ballard, D. H., Hayhoe, M. M., Pook, P. K., & Rao, R. P. N. (1997). Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences*, *20*, 723–767.

Barr, D. J. (2008). Analyzing 'visual world' eyetracking data using multilevel logistic regression. *Journal of memory and language*, *59*(4), 457–474.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random-effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255-278.

Castelhano, M. S., & Henderson, J. M. (2007). Initial scene representations facilitate eye movement guidance in visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *33*, 753–763.

Cavanagh, P., & Alvarez, G. A. (2005). Tracking multiple targets with multifocal attention. *Trends in Cognitive Sciences*, *8*(7), 349–354.

Chen, Q., & Mirman, D. (in press). Interaction between phonological and semantic representations: Time matters. *Cognitive Science*.

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

Coco, M. I., & Keller, F. (2015a). Integrating mechanisms of visual guidance in naturalistic language production. *Cognitive processing*, *16*(2), 131–150.

Coco, M. I., & Keller, F. (2015b). The interaction of visual and linguistic saliency during syntactic ambiguity resolution. *The Quarterly Journal of Experimental Psychology*, *68*(1), 46–74.

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

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, *8*, 1117–1121.

Demberg, V., Keller, F., & Koller, A. (2013). Incremental, predictive parsing with psycholinguistically motivated tree-adjoining grammar. *Computational Linguistics*, *39*(4). (1025–1066)

Eckstein, M. P., Drescher, B. A., & Shimozaki, S. S. (2006). Attentional cues in real scenes, saccadic targeting, and bayesian priors. *Psychological Science*, *17*(11), 973–980.

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

Federmeier, K. D. (2007). Thinking ahead: The role and roots of prediction in language comprehension. *Psychophysiology*, *44*(4), 491–505.

Federmeier, K. D., & Kutas, M. (1999). A rose by any other name: Long-term memory structure and sentence processing. *Journal of Memory and Language*, *41*, 469–495.

Federmeier, K. D., Wlotko, E., De Ochoa-Dewald, E., & Kutas, M. (2007). Multiple effects of sentential constraint on word processing. *Brain Research*, *1146*, 75–84.

Ferreira, F., Apel, J., & Henderson, J. M. (2008). Taking a new look at looking at nothing. *Trends in Cognitive Science*, *12*(11), 405–410.

Friesen, C. K., & Kingstone, A. (1998). The eyes have it! reflexive orienting is triggered by nonpredictive gaze. *Psychonomic Bulletin & Review*, *5*, 490–495.

Greene, M. R., & Oliva, A. (2009). Recognition of natural scenes from global properties: Seeing the forest without representing the trees. *Cognitive psychology*, *58*(2), 137–176.

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

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

Hoover, M. A., & Richardson, D. C. (2008). When facts go down the rabbit hole: Contrasting features and objecthood as indexes to memory. *Cognition*(108), 533–542.

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.

Huettig, F., Quinlan, P. T., McDonald, S. A., & Altmann, G. T. M. (2006). Models of high-dimensional semantic space predict language-mediated eye movements in the visual world. *Acta Psychologica*, *121*(1), 65–80.

Hwang, A. D., Wang, H.-C., & Pomplun, M. (2011). Semantic guidance of eye movements in real-world scenes. *Vision Research*, *51*(10), 1192–205.

Kahneman, D., Treisman, A., & Gibbs, B. J. (1992). The reviewing of object files: Object-specific integration of information. *Cognitive Psychology*, *24*(2), 175–219.

Kamide, Y., Altmann, G. T. M., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language*,

49, 133 - 156.

Kamide, Y., Scheepers, C., & Altmann, G. T. M. (2003). Integration of syntactic and semantic information in predictive processing: Cross-linguistic evidence from German and English. *Journal of Psycholinguistic Research*, *32*, 37–55.

Knoeferle, P., & Crocker, M. (2006). The coordinated interplay of scene, utterance and world knowledge. *Cognitive Science*, *30*, 481–529.

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

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, *104*(2), 211–240.

Lau, E., Holcomb, P. J., & Kuperberg, G. R. (2013). Dissociating N400 effects of prediction from association in single word contexts. *Journal of Cognitive Neuroscience*, *25*(3), 484–502.

Levy, R., & Keller, F. (2013). Expectation and locality effects in German verb-final structures. *Journal of Memory and Language*, *68*(2), 199–222.

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

Masson, M. E. J., & Loftus, G. R. (2003). Using confidence intervals for graphically based data interpretation. *Canadian Journal of Experimental Psychology*, *57*, 203–220.

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.

Mirman, D., & Magnuson, J. (2009). Dynamics of activation of semantically similar concepts during spoken word recognition. *Memory & Cognition*, *37*(7), 1026-1039.

Neider, B., M., & Zelinsky, G. (2006). Scene context guides eye movements during visual search. *Vision Research*, *46*, 614–621.

Pinheiro, J., & Bates, D. (2000). *Mixed-effects models in s and s-plus*. Springer-Verlag.

Pylyshyn, Z. W. (1989). The role of location indexes in spatial perception: a sketch of the FINST spatial index model. *Cognition*, *32*, 65–97.

Richardson, D. C., Altmann, G. T. M., Spivey, M. J., & Hoover, M. A. (2009). Much ado about eye movements to nothing: a response to ferreira et al.: taking a new look at looking at nothing. *Trends in Cognitive Science*, *13*(6), 235–236.

Richardson, D. C., & Spivey, M. J. (2000). Representation, space and hollywood squares: looking at things that aren't there anymore. *Cognition*, *76*, 269–295.

Rosenholtz, R., Li, Y., & Nakano, L. (2007). Measuring visual clutter. *Journal of Vision*, *7*(2), 1–22.

Russell, B., Torralba, A., Murphy, K., & Freeman, W. (2008). Labelme: a database and web-based tool for image annotation. *International journal of computer vision*, *77*(1-3), 157–173.

Spivey, M., & Geng, J. (2001). Oculomotor mechanisms activated by imagery and memory: Eye movements to absent objects. *Psychological Research*, *65*, 235–241.

Staub, A., Abbott, M., & Bogartz, R. S. (2012). Linguistically-guided anticipatory eye movements in scene viewing. *Visual Cognition*, *20*, 922–946.

Staub, A., & Clifton, C. (2006). Syntactic prediction in language comprehension: Evidence from either ... or. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*, 425-436.

Staudte, M., & Crocker, M. (2011). Investigating joint attention mechanisms through spoken human-robot interaction. *Cognition*, *120*, 268–291.

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

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.

van Berkum, J. J. A., Brown, C. M., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating upcoming words in discourse: Evidence from erps and reading times. *Journal of Experimental Psychology:*

Learning, Memory, and Cognition, *31*, 443-467.

Yee, E., & Sedivy, J. (2006). Eye movements to pictures reveal transient semantic activation during spoken word recognition. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *32*(1), 1–14.

Appendix
Sentence Materials

This appendix contains the sentences used in Experiments 1 and 2. We also give the context words in square brackets. These were used for the predictivity and plausibility norming study. Note that fort the plausibility norming study a preposition was inserted before the context words if required to avoid grossly implausible items (e.g., *the man ate at the table* rather than *the man ate the table*).

(5) a. The woman sailed the yacht. [ocean]
b. The woman viewed the yacht. [ocean]

(6) a. The man putted the ball. [course]
b. The man detested the ball. [course]

(7) a. The woman parked the car. [driveway]
b. The woman cleaned the car. [driveway]

(8) a. The woman smelled the flowers. [grass]
b. The woman discounted the flowers. [grass]

(9) a. The man rang the doorbell. [wall]
b. The man heard the doorbell. [wall]

(10) a. The man burned the hamburgers. [grill]
b. The man moved the hamburgers. [grill]

(11) a. The woman rinsed the dishes. [sink]
b. The woman checked the dishes. [sink]

(12) a. The woman presented the slide. [screen]
b. The woman ignored the slide. [screen]

(13) a. The man lit the fire. [fireplace]
b. The man photographed the fire. [fireplace]

(14) a. The woman rebooted the laptop. [desk]
b. The woman inspected the laptop. [desk]

(15) a. The man dribbled the ball. [court]
b. The man watched the ball. [court]

(16) a. The man raised the flag. [pole]
b. The man perceived the flag. [pole]

(17) a. The woman dumped the bottle. [trash]
b. The woman examined the bottle. [trash]

(18) a. The woman rowed the boat. [water]
b. The woman held the boat. [water]

(19) a. The man dealt the cards. [table]
b. The man noticed the cards. [table]

(20) a. The boy fed the fish. [aquarium]
b. The boy liked the fish. [aquarium]

(21) a. The woman hung the painting. [wall]

b. The woman saw the painting. [wall]

(22) a. The woman browsed the books. [shelf]
b. The woman forgot the books. [shelf]

(23) a. The man hailed the taxi. [road]
b. The man observed the taxi. [road]

(24) a. The woman painted the picture. [easel]
b. The woman shifted the picture. [easel]

(25) a. The man ate the sandwich. [table]
b. The man removed the sandwich. [table]

(26) a. The woman milked the cow. [grass]
b. The woman bought the cow. [grass]

(27) a. The woman landed the plane. [runway]
b. The woman loved the plane. [runway]

(28) a. The man docked the boat. [sea]
b. The man borrowed the boat. [sea]