Visual context constrains language-mediated anticipatory eye movements

Florian Hintz,[1] Antje S. Meyer, [1,2] and Falk Huettig[1,2]

[1]*Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands*

[2]*Radboud University, Nijmegen, The Netherlands*

---In press at Quarterly Journal of Experimental Psychology---

Running title: Language prediction and visual context

Keywords: Predictive language processing; eye movements; visually-induced competition

Correspondence should be addressed to:

Florian Hintz

Max Planck Institute for Psycholinguistics

P.O. Box 310

6500 AH Nijmegen

The Netherlands

Telephone: +31 24 3521335

Florian.Hintz@mpi.nl

Abstract

Contemporary accounts of anticipatory language processing assume that individuals predict upcoming information at multiple levels of representation. Research investigating language-mediated anticipatory eye gaze typically assumes that linguistic input restricts the domain of subsequent reference (visual target objects). Here, we explored the converse case: Can visual input restrict the dynamics of anticipatory language processing? To this end, we recorded participants' eye movements as they listened to sentences in which an object was predictable based on the verb's selectional restrictions ("The man peels a banana"). While listening, participants looked at different types of displays: The target object (banana) was either present or it was absent. On target-absent trials, the displays featured objects that had a similar visual shape as the target object (canoe) or objects that were semantically related to the concepts invoked by the target (monkey). Each trial was presented in a long preview version, where participants saw the displays for approximately 1.78 seconds *before* the verb was heard (pre-verb condition), and a short preview version, where participants saw the display approximately 1 second *after* the verb had been heard (post-verb condition), 750 ms prior to the spoken target onset. Participants anticipated the target objects in both conditions. Importantly, robust evidence for predictive looks to objects related to the (absent) target objects in visual shape and semantics was found in the post-verb but not in the pre-verb condition. These results suggest that visual information can restrict language-mediated anticipatory gaze and delineate theoretical accounts of predictive processing in the visual world.

41 **Visual context constrains language-mediated anticipatory eye movements**

42

43 ## Introduction

44 There is now broad consensus that people often predict which words will come next (e.g., Altmann
45 & Mirković, 2009; Dell & Chang, 2014; Federmeier, 2007; Huettig, 2015; Kamide, 2008; Kutas,
46 DeLong, & Smith, 2011; Pickering & Garrod, 2013), for example, when reading books or having a
47 conversation about politics or the state of the environment. In these cases, people may primarily rely
48 on linguistic processes when anticipating the continuation of sentences that do not refer directly to
49 something happening in their visual surroundings. However, there are many situations in everyday
50 conversations where speakers refer to objects or actions in their immediate environment. In these
51 circumstances, listeners must integrate linguistic input with relevant real world visual context (e.g.,
52 Henderson & Ferreira, 2004).

53 Several studies have shown that visual context directly affects eye gaze related to language
54 processing (e.g., Altmann & Kamide, 2007; Chambers et al., 2002; Coco, Keller, & Malcolm, 2016;
55 Ferreira, Foucart, & Engelhardt, 2013; Knoeferle & Crocker, 2006; Tanenhaus, Spivey-Knowlton,
56 Eberhard, & Sedivy, 1995) but surprisingly little is known about the exact nature of such
57 interactions. Although the assumption of strong modularity between language and vision systems
58 (Fodor, 1993) has long been discredited (e.g., Anderson, Chiu, Huette, & Spivey, 2011), the extent of
59 interactivity between the two information-processing streams is still unclear. One possibility is that
60 when anticipating what a speaker may say next listeners' visual and linguistic signals converge on a
61 common processing stream (e.g. Altmann & Mirković, 2009; cf. Prinz, 1997). According to such a
62 common coding approach, visual and linguistic processing share a common representational
63 substrate (i.e. visual and linguistic events are represented in the same way). Another possibility is
64 that linguistic and visual (scene) processing are tightly interrelated but partly independent (e.g., the
65 coordinated interplay account, Knoeferle & Crocker, 2006). According to this view, utterance

meaning, linguistic expectations and visual context are related through processes of co-indexation and subsequent resolution of linguistic and visual (scene) processing. In the present study, we tested these proposals using the visual world eye-tracking method that has been instrumental for formulating both the common coding (Altmann & Mirković, 2009) and the coordinated interplay (Knoeferle & Crocker, 2006) accounts.

The common coding account relates to previous work by Altmann and Kamide (1999). They presented participants with semi-realistic drawings depicting for instance a boy, a cake, and a number of inedible objects. While participants were viewing these scenes, they heard sentences such as "The boy will eat the cake" or "The boy will move the cake". Eye movement recordings suggested that participants anticipated 'cake' (referring to the only edible object in the scene) on hearing "eat" but not on hearing "move". Altmann and Kamide (2007) explained this anticipation effect in the same way as they interpreted semantic competition effects in the visual world paradigm (i.e. participants fixating semantically related objects such as a trumpet on hearing "piano", Huettig & Altmann, 2005; cf. Yee & Sedivy, 2006). That is, according to their account, the conceptual overlap between the mental representation accessed when hearing "piano" and the representation previously activated from seeing the trumpet boosts the activation of the representation corresponding to the trumpet, and this results in increased likelihood of executing a saccadic eye movement toward it. Similarly, Altmann and Mirković (2009) argued that the activation of 'cake' is boosted on hearing "eat", because linguistic and visual processing are subserved by "the same underlying process", that "manifest[s] across the same representational substrate" (Altmann & Mirković, 2009, p. 601).

However, some visual world data are inconsistent with such an account. Huettig and Altmann (2007) asked participants to listen to sentences such as "In the beginning, the zookeeper worried greatly but then he looked at the snake and realized it was harmless", while seeing a display featuring clip art pictures of the target (snake) and three unrelated distractor objects. The authors found that participants already looked longer at the picture of the snake than at the distractor pictures when

hearing “zookeeper”, indicating that they anticipated the snake to be referred to. In a different condition, when the target object was replaced with an object that had the same visual shape as the target (electric cable), Huettig and Altmann observed more looks to visual competitors than to the unrelated distractors when participants heard the word “snake”. This behaviour suggests that participants integrated the visual shape information retrieved from the displays with visual shape information that became available as the spoken word unfolded (cf. Dahan & Tanenhaus, 2005; Huettig & Altmann, 2004). Importantly, no such bias to the visual competitors was observed earlier, before the onset of “snake”, when participants heard “zookeeper”. These results are inconsistent with a strong interpretation of the common coding account, because it assumes that the zookeeper context should activate ‘snake’ and its visual shape, which overlaps with ‘cable’ (already strongly active in the common representational substrate because of its presence in the visual scene). Thus, according to the common coding account, the likelihood of a saccade towards the picture of the cable should be higher than to the unrelated distractors (unless an additional inhibitory mechanism is postulated, which suppresses looks to the competitor object).

Interestingly, a subsequent study found experimental evidence for anticipatory looks to visual competitors (Rommers et al., 2013). A crucial difference between Huettig and Altmann (2007) and Rommers and colleagues (2013) was that in the latter study participants were given only a short preview of the visual display, starting 500 ms before the onset of the target word, while participants in Huettig and Altmann’s study had ample time to inspect the display (5 seconds prior to target word onset). In sum, these results suggest that (a) listeners’ predictions about upcoming words can include visual shape information about that concept and that (b) the available preview time might affect anticipatory language-mediated eye movements.

Indeed, a previous study by Huettig and McQueen (2007) has shown that preview time has strong effects on the likelihood of word-object mapping at different representational levels. The participants in that study heard sentences such as “Eventually, she looked at the beaker that was

116 standing in front of her", where beaker was the target word. Coinciding with the onset of the spoken
117 sentence, participants were presented with displays containing an object that was semantically
118 related to the target word (e.g. fork), an object that had the same visual shape as the target (e.g.
119 bobbin), an object that overlapped with the target word in phonological onset (e.g. beaver), and an
120 unrelated distractor (e.g. umbrella). The target word was preceded by on average seven words,
121 providing participants with ample time to preview the display. Analysing eye movements from the
122 onset of the target words, Huettig and McQueen observed a pattern, which they termed the *tug of*
123 *war* between phonological, semantic and visual shape information: About 200 ms after target word
124 onset, participants looked at the picture of the beaver due to the phonological overlap for as long as
125 "beaker" and "beaver" overlapped and only later looked more at the pictures of the fork and the
126 bobbin, demonstrating matches in semantics and visual shape, respectively. Explaining the complex
127 interplay between vision and language systems, the authors reasoned that when previewing the
128 displayed objects, activation in the visual processing system cascaded from visual levels to semantic
129 levels to phonological processing levels (i.e. the object name) and that in the linguistic system,
130 activation cascaded from phonological (i.e. hearing the object's name) to semantic and visual levels.
131 In their Experiment 2, Huettig and McQueen reduced the display preview time to 200 ms prior to
132 target word onset and found that participants' fixation behaviour was determined by matches at
133 semantic and visual levels only, suggesting that by the time of target word onset, activation had not
134 cascaded to phonological levels yet. Taken together, preview time and the coordination of visual and
135 linguistic processing streams more generally appear to be crucial for determining the levels at which
136 word-object mappings take place. Arguably, these types of vision-on-language effects can be
137 accommodated more straightforwardly by an account, where linguistic and visual (scene) processing
138 are tightly interrelated but partly independent. In their coordinated interplay account, Knoeferle and

Crocker (2006), for example, propose that the attended visual context rapidly influences linguistic comprehension of the concurrent speech[1].

Could the difference in preview time have affected the likelihood of looks to visual competitors in Huettig and Altmann's (2007) and Rommers and colleagues' (2013) studies and thus explain the different patterns of results? We addressed this question in the present study. In the present experiment, which is in many ways similar to Huettig and Altmann's and Rommers et al.'s studies, we varied the timing of the presentation of the sentence relative to the relevant display within-participants. Specifically, participants listened to sentences where the final word was predictable based on verb thematic role assignment (e.g., Dutch translation equivalent of "The man peels at that moment a banana"). While hearing the spoken sentences, they looked at displays showing four clip art pictures. On target-present trials, one of them was the target (banana), and the other objects were unrelated distractors. On target-absent trials, the target was replaced with a visually similar object (canoe) or a semantically related object (monkey). Crucially, presentation of the display began either 1.78 seconds before the verb was heard (pre-verb condition), or 750 ms before the onset of the target, which was 1 second after the verb had been heard (post-verb condition).

In the post-verb condition, i.e. when linguistic processing (hearing the verb "peel") precedes viewing the visual display, one would expect more looks to the target (when present) than to the unrelated distractors. One would also expect more looks to the semantic and visual competitors than to the distractors. Importantly, this behaviour is predicted by both common coding and coordinated interplay accounts. The common coding account predicts looks to the competitors, because linguistic and visual processing of semantic and visual shape information are assumed to converge on the same representational substrate. The coordinated interplay account predicts looks to the semantic and visual competitors because cascaded processing in the visual processing stream has not yet advanced

[1] See Ferreira et al. (2013) and Coco et al. (2016) for related accounts.

to higher (e.g. phonological; cf. Huettig & McQueen, 2007, Experiment 2) levels. Thus, a match between visually-derived and linguistically-derived semantic and visual representations occurs triggering an eye movement to the competitor objects.

As both accounts predict anticipatory semantic and visual shape effects, we view the post-verb condition as a baseline to establish the size of the competition effects and to demonstrate the suitability of the selected materials. The crucial question was how the gaze patterns in the pre-verb condition would compare to those in the post-verb condition. As in the post-verb condition, one would expect more looks to the target than to the unrelated distractors in the target-present condition replicating Altmann and Kamide's (1999) seminal study. The predictions for the semantic and visual competitors depend on the nature of the interaction between language and vision systems. Assuming that visual and linguistic processing converge on a common representational substrate (i.e. visual and linguistic information are processed in the same cognitive space, common coding account), preview time of the visual objects should not substantially modulate semantic and visual competition (as compared to the post-verb condition) and we should also observe anticipatory semantic and visual competition effects in the pre-verb condition. In other words, viewing the competitor scenes (including a monkey and a canoe) should result in the activation of object representations (e.g., semantic and visual features) preceding the arrival of language. Hearing a sentence that leads to the prediction of a semantically or visually related object (banana) should increase the activation level of the competitor objects (compared to the unrelated distractors) and should result in anticipatory eye gaze, or as Altmann & Mirković, 2009, p. 593, put it: "overlapping components increase in activation because they receive dual support" from visual and spoken input

On an account, where listeners' visual processing and linguistic processing comprise two tightly related but separate streams (coordinated interplay account), the competition effects in the pre-verb and post-verb conditions should differ. This is because processing of the visual objects during preview and processing of the spoken input lead to activation of partly independent,

separately represented information: Given the long preview time, activation of object information in the visual processing stream should cascade from visual to semantic and finally to phonological levels of representation (i.e. the object labels for monkey and canoe; cf. Huettig & McQueen, 2007). In the linguistic processing stream, hearing "peel" is assumed to activate 'banana' (including semantic and visual information about it). Critically, as listeners have retrieved labels for the displayed objects, phonological information should dominate language-mediated visual search, thereby constraining anticipatory eye movements to the extent that the biases towards the visual and semantic competitors should be reduced or absent in the pre-verb condition (cf. Huettig & Altmann, 2007).

Method

Participants

Sixty members of the subject panel of the MPI (eleven males, mean age = 22, *SD* = 3), took part in the experiment. All were native speakers of Dutch and did not report any history of learning or reading disabilities or neurological or psychiatric disorders. The participants were paid for participation. The ethics board of the Faculty of Social Sciences at Radboud University approved the study.

Materials

The materials consisted of 30 Dutch transitive sentences (e.g., "De man pelt op dit moment een banaan", the man peels at that moment a banana) in which the final word was predictable based on the selectional restrictions of the verb. All sentences had the same structure and the same number of words: The subject position was taken by "the man", and the adverbial "at that moment" separated verb and target to ensure that participants had enough time to generate predictions and to program and launch saccadic eye movements prior to the onset of the spoken targets. The resulting sentence

213 construction is deemed to be quite natural by native Dutch speakers. The mean word frequency of
214 the target nouns was 25 per million words (Keuleers et al., 2010; *SD* = 30); the mean frequency of
215 the inflected verbs was 4 per million (*SD* = 7; six verbs were not listed).
216 The sentences were pre-tested for cloze probability (Taylor, 1953) using an online tool for
217 web experiments developed by the technical group of the MPI. Thirty-eight Dutch native speakers
218 (five males; mean age = 22; *SD* = 3) took part in the rating study; none of whom participated in other
219 rating studies or the main experiment. The participants read the sentences until the object position
220 and were asked to fill in the word that would in their opinion best complete the sentence. The cloze
221 probability of a sentence was defined as the proportion of the word in question divided by all
222 responses provided for that sentence. The average cloze probability for all sentences was .23 (*SD* =
223 .25).
224 To create the visual displays, we used the stimulus set by de Groot et al. (2016), which
225 contains words and photographs of common objects matched for visual and semantic similarity. For
226 each of the 30 targets, we selected a visual competitor, i.e., an object that had a similar visual shape
227 as the concept invoked by the target and a semantic competitor, i.e., an object that was semantically
228 similar to the target. Both competitors were unrelated to the target on all other dimensions (de Groot
229 et al., 2016, for details of the ratings procedure and definition of visual and semantic similarity). The
230 target fulfilled the selectional restrictions of the verb far better than the competitors did. We
231 controlled that the association strength between the sentence verb and the two competitors was zero,
232 using a Dutch free association database (De Deyne & Storms 2008). For each target, we also selected
233 three unrelated objects as distractors[2] and a picture of the target. All pictures had the same size and
234 resolution (124 x 124 pixels, 72 dpi).

[2] As de Groot et al.'s (2016) stimulus set only provides norms for two unrelated distractor objects per target word, we carried out additional semantic similarity and visual similarity rating studies (n = 36, nine males, mean age = 22, *SD* = 3, none of these volunteers took part in the main experiment or the cloze probability rating study) on the third distractor following de Groot et al.'s procedure. The additional distractors were rated not to be visually or semantically similar to the concept invoked by the target noun (visual rating task: 1.55; *SD* = 1.54; semantic rating task: .45; *SD* = .59; on a 1-10 scale).

Each sentence was paired with three different displays each showing three unrelated distractors and one of three critical objects (e.g., target, semantic competitor, visual competitor, Figure 1, for a schematic of an experimental display). Each display type was presented in a pre-verb and post-verb trial. The six versions of each item were distributed across six experimental lists such that each sentence occurred only once on one list. The lists featured equal numbers of pre-verb and post-verb trials. Each display type occurred ten times on each list. In order to create equal numbers of target-present and target-absent trials, we added ten filler sentences, which had the same structure as the experimental sentences and which were low in predictability (cloze probability, assessed in the same rating study as described above, was zero). The filler sentences were paired with displays containing a picture of the target and three unrelated distractors and also occurred as pre-verb and post-verb version.

Figure 1

Procedure

The 30 experimental sentences and the ten filler sentences were spoken with neutral intonation at a normal pace by a female native speaker of Dutch. Recordings were made in a sound-damped booth, sampling at 44 kHz (mono, 16-bit sampling resolution). The mean duration of the experimental sentences was 3231 ms (*SD* = 195). Onsets and offsets of all words were marked using Praat (Boersma, 2002). The time between the onset of the verb and the onset of the target noun in the experimental sentences was on average 1810 ms (*SD* = 184). In these sentences, the average duration of the target nouns was 608 ms (*SD* = 111); the average duration of the inflected verbs was 555 ms (*SD* = 112).

The participants were tested individually in a sound-shielded booth. Eye movements were recorded using an EyeLink 1000 tracker sampling at 1000 Hz. Participants placed their heads in a

chinrest, which was approximately 75 cm away from the computer screen. The experimental stimuli were shown on a 23-inch computer screen, in a region spanning 1024 x 768 pixels. After calibration, participants were randomly assigned one list. The order of trials was random with the constraint that maximally two trials of the same display type appeared in a row. The spoken sentences were presented through headphones. A trial started with the presentation of a central fixation dot for two seconds. On pre-verb trials, the dot was replaced with the display and the playback of the sentence started after one second. In the experimental sentences, the onset of the verbs occurred on average after 784 ms (*SD* = 112), amounting to approximately 1.78 seconds of visual preview before the verb was heard. On post-verb trials, the playback of the spoken sentence started immediately after the two-second presentation of the fixation dot; the presentation of the displays was timed to begin 750 ms before the onset of the spoken target, which was approximately 1 second after the spoken onset of the verb. All objects had the same distance from the centre, with a direct visual angle of about 12°. In both preview conditions, the four objects remained in view until the end of the trial (see Figure 2, for a schematic of the trial structure). The positions of the four objects were randomized. The participants carried out a look-and-listen task (Huettig et al., 2011a, for discussion), which means that they should listen carefully and could look at whatever they wanted while not moving their eyes away from the computer screen.

Regions of interests (250 x 250 pixels) were defined around each object. The data from participants' left or right eye (depending on the quality of the calibration) were analysed in terms of fixations, saccades, and blinks by the algorithm provided in the EyeLink software. Fixations on experimental trials were coded as directed to the target, semantic competitor, visual competitor, one of the three distractors, or elsewhere.

Figure 2

Results

Due to track loss, a total of 21 out of 1800 experimental trials had to be removed. Figure 3 shows participants' eye movements in both preview conditions and the three display types for a time window starting 2000 ms before the onset of the spoken target until 1000 ms post target onset. By-participant confidence intervals (95%), computed at each sampling step (1 ms), were added to all lines indicating by-participant variance (Masson & Loftus, 2003; cf. Fidler & Loftus, 2009). The area between the lower and the upper bounds is shaded in grey. Note again that on post-verb preview trials, participants were fixating a dot in the centre of the screen until the visual display was presented 750 ms before spoken target onset. This yielded fixation proportions around zero for a large part of the trial. Visual inspection of Figure 3 suggests that fixations were first directed to any of the objects in the post-verb condition around 250 ms after presentation of the display. This is because takes minimally 200 ms to program and launch a language-mediated saccadic eye movement (cf. Saslow, 1967).

Figure 3+Table 1

The top panels in Figure 3 show that participants anticipated the targets on both pre-verb and post-verb trials. On post-verb trials, anticipatory eye movements to the target objects arose around 500 ms before the objects were referred to in the speech signal. On pre-verb trials, participants gazed at the target objects shortly after having recognized the verbs, around one second prior to the target onset. The middle panels show that on post-verb trials participants showed a strong bias towards the semantic competitor. On pre-verb trials, we observed a tendency for a bias in looks to the semantic competitor, which arose shortly before the spoken target was heard. The bottom panels show that there was a bias towards the visual competitors on post-verb trials but not pre-verb trials. Finally, for both preview types, we observed fixations to visual and semantic competitors at around 500 ms after

the target onset, which most likely reflect bottom-up processing of the spoken target (cf. Dahan & Tanenhaus, 2005; Duñabeitia et al., 2009; Huettig & Altman, 2005, 2007; Yee & Sedivy, 2006).

To analyse the differences between pre-verb and post-verb conditions statistically (see Table 1, for mean fixation proportions), we fitted a linear mixed-effects model in R (R Development Core Team, 2012) using the lme4 package (Bates et al., 2015). For both preview conditions, the dependent variable was calculated for the period starting 500 ms before target word onset (when the first fixation was made to any of the objects in the post-verb condition) and ending 200 ms after target word onset. Fixation proportions were transformed to log odds, the appropriate scale for assessing effects on a categorical dependent variable, using the empirical logit function (Barr, 2008). The average log odds of looks to the three unrelated distractors was subtracted from the average log odds of looks to the target/semantic competitor/visual shape competitor object to create the dependent variable, which indicates the strength of any bias toward each experimental picture over the unrelated distractor pictures. The model contained *preview* (pre-verb vs. post-verb) and *display type* (target vs. semantic competitor vs. visual shape competitor) as fixed factors and the interaction of both. Both factors were treatment-coded. *Participant* and *item* were added as random effects. The random effects structure further contained random intercepts and random slopes for *preview* by *participant* and *item* (Baayen, Davidson, & Bates, 2008; more complex models, containing random slopes for *display type*, failed to converge). This 'maximal' model was compared, using the anova()-command, to a model that was identical in random effects structure to the maximal model but did not contain the interaction between *preview* and *display type* factors. Dropping the interaction led to significantly worse model fit ($\chi(2) = 6.73$, $p = 0.035$). The final model formula was thus *empirical_log ~ preview * display type + (1+preview | participant) + (1+preview | item).* The *pre-verb preview* condition and the *target-present display* condition were put on the intercept. P-values were obtained using the lmerTest package (version 2.0-33, Satterthwaite degrees of freedom

approximation, Kuznetsova, Brockhoff, & Christensen, 2016). Post-hoc contrasts were performed using emmeans (Kenward-Roger's approximation to degrees of freedom, Lenth, 2018).

Table 2

The model revealed a simple effect of preview type (Table 2, for an overview; $\beta = 1.85$, SE = 0.52, $t = 3.6$, $p < 0.001$) with stronger biases for the critical objects on post-verb than on pre-verb trials. The post-hoc contrasts showed that all critical objects were looked at more in the post-verb than in the pre-verb condition: pre-verb target vs. post-verb target: $\beta = -1.85$, SE = 0.52, $t = -3.6$, $p < 0.001$; pre-verb semantic competitor vs. post-verb semantic competitor: $\beta = -2.43$, SE = 0.52, $t = -4.7$, $p < 0.001$; pre-verb visual competitor vs. post-verb visual competitor: $\beta = -1.07$, SE = 0.52, $t = -2.07$, $p = 0.042$.

Based on a reviewer suggestion, we also added cloze probability (scaled and centred) as a continuous predictor to the mixed-effects model described above (formula: *empirical_log ~ preview * display type + cp + (1+preview | participant) + (1+preview | item)*. The contribution of cloze probability to explaining variance in the dependent variable was minimal ($\beta = -0.04$, SE = 0.19, $t = -0.2$, $p > 0.1$). Moreover, having cloze probability in the model did not affect the main results (pre-verb target vs. post-verb target: $\beta = -1.85$, SE = 0.52, $t = -3.6$, $p < 0.001$; pre-verb semantic competitor vs. post-verb semantic competitor: $\beta = -2.43$, SE = 0.52, $t = -4.7$, $p < 0.001$; pre-verb visual competitor vs. post-verb visual competitor: $\beta = -1.07$, SE = 0.52, $t = -2.01$, $p = 0.042$).

Discussion

We investigated the influence of timing of the availability of visual input on the likelihood of anticipatory eye movements to objects semantically and visually related to predicted target objects. To that end, we manipulated the time participants received to preview the visual displays. In the target-present condition, we observed anticipatory eye movements to objects that satisfied the

thematic role requirements of the verb with pre-verb and post-verb preview manipulations. This replicates previous research showing that listeners anticipate upcoming nouns/visual referents (e.g., Altmann & Kamide, 1999). On post-verb trials, we found a strong semantic and a weaker visual shape bias (replicating Rommers et al., 2013). These effects were eliminated, or strongly reduced, in the pre-verb condition.

The differences between pre-verb and post-verb conditions on semantic and visual competitor trials are inconsistent with a strong interpretation of a common coding account of language-vision interactions, where visual and linguistic signals converge on a common representational substrate (e.g., Altmann & Mirković, 2009). Such an account would predict similar behaviour in both preview conditions because visual processing (seeing the pictures of a monkey and a canoe) and linguistic processing (hearing a sentence biasing towards banana) should both increase the activation level of the competitor objects and thereby the likelihood of eye movements towards them. The present results may, however, be compatible with a common coding account if additional mechanisms, such as inhibition of related representations are postulated. Future research could be conducted to explore this possibility.

Our results fit more straightforwardly with the view that language-mediated anticipatory eye movements are subserved by separate, but tightly interacting visual and linguistic processing streams (e.g., Knoeferle & Crocker, 2006). Specifically, one interpretation of the present data is that extensive visual preview leads to the retrieval of linguistic information (e.g. phonological information; Huettig & McQueen, 2007; Mani & Plunkett, 2010; McQueen & Huettig, 2014, for further experimental evidence) from the visual processing stream, making ‘object labels’ available for word-object mappings. In line with such an interpretation, one might conjecture that with extensive preview anticipatory word-object mappings primarily take place at phonological levels of representation. By contrast, with short preview, word-object mappings occur at semantic and visual

levels as well. This interpretation of the data resonates with previous research (e.g., Lupyan, 2012) highlighting the importance of object labels for cognitive processing.

Additionally, on pre-verb trials participants had ample time to look at the displays and thus knew which objects were and which ones were not present when they heard the sentence. On post-verb trials, on the other hand, preview time was much reduced, which may have resulted in a greater likelihood of attentional capture or 'pop-out' effects (cf. Yantis & Jonides, 1984) by related objects (i.e. semantic and visual competitors) than on pre-verb trials. Indeed, the strong semantic competitor bias on post-verb trials suggests that while participants were predicting the target object, semantic competitors captured their attention shortly after display onset and they continued to look at them for an extended period of time (i.e. 400 ms after target word onset).

This is not to say that competition effects, i.e. looks to semantically and visually related objects, do not occur with substantial preview periods. For example, the participants in the study by Huettig and Altmann (2005; see also Dahan & Tanenhaus, 2005; Duñabeitia et al., 2009; Yee & Sedivy, 2006) received a one-second preview of the visual scene before the playback of the spoken sentence containing the target word. That means they had plenty of time to inspect the objects and most likely knew which objects were in the scene when the spoken sentence commenced. Crucially, shortly after the onset of the target word (e.g., "trumpet", not present in the scene) participants started to look at the picture of the semantic competitor (e.g. piano). Thus, a competitor effect occurred in spite of the long preview period. Future research could investigate the interaction between preview time and attentional capture more thoroughly.

The present pattern of results are also in line with a recent account by Coco and colleagues (2016; cf. Altmann & Kamide, 2007, 2009), who argued that the visual scene provides contextual guidance for language processing. Coco and colleagues emphasized that the usage of real-world photographs was crucial for seeing scene-specific effects of vision on anticipatory language processing, 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" (p. 22). Though using more naturalistic visual stimuli may increase the likelihood of vision-on-language effects, the present study shows that it is not a prerequisite (cf. Saryazdi & Chambers, 2018). Our experimental setup, using incoherent scenes featuring four distinct unrelated visual objects, enabled us to determine that knowledge retrieved from viewing visual objects can constrain linguistic prediction even in the absence of a coherent visual scene.

To conclude, adding to a growing body of data, we provide experimental evidence showing that preview time impacts language-mediated anticipatory gaze: When speech is accompanied by relevant visual context, listeners' eye movements to upcoming referents are constrained by information extracted from the visual context. Specifically, we believe that the present data are most compatible with a view where listeners exploit the preview phase to retrieve phonological information about co-present visual input, which constrains anticipatory looks to objects partially matching the predicted target. Such an interpretation is inconsistent with the notion that language-mediated anticipatory eye movements are subserved by a common coding system where linguistic and visual processing converge on a single substrate. Instead, we endorse the view that linguistic and visual processing comprise separate streams that interact tightly. Future research is needed to corroborate this claim and to rule out alternatives.

References

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: restricting the domain of subsequent reference. *Cognition, 73*(3), 247-264.

Altmann, G. T., & 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*(4), 502-518.

Altmann, G. T., & Kamide, Y. (2009). Discourse-mediation of the mapping between language and the visual world: Eye movements and mental representation. *Cognition*, *111*(1), 55-71.

Altmann, G. T. M., & Mirkovic, J. (2009). Incrementality and Prediction in Human Sentence Processing. *Cognitive Science, 33*(4), 583-609.

Anderson, S. E., Chiu, E., Huette, S., & Spivey, M. J. (2011). On the temporal dynamics of language-mediated vision and vision-mediated language. *Acta psychologica*, *137*(2), 181-189.

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

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

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67, 48.

Boersma, P. P. G. (2002). Praat, a system for doing phonetics by computer (Version 5.1.19) [Computer program]. *Available from http://www.praat.org/*.

Chambers, C.G., Tanenhaus, M.K., Eberhard, K.M., Carlson, G.N. & Filip, H. (2002). Circumscribing referential domains during real time language comprehension. *Journal of Memory and Language, 47*, 30-49.

Coco, M. I., Keller, F. & Malcolm, G.L. (2016). Anticipation in Real-World Scenes: The Role of Visual Context and Visual Memory. *Cognitive Science 40*(8), 1995-2024.

De Groot, F., Koelewijn, T., Huettig, F., & Olivers, C. N. L. (2016). A stimulus set of words and pictures matched for visual and semantic similarity. *Journal of Cognitive Psychology, 28*(1), 1-15.

Dahan, D., & Tanenhaus, M. K. (2005). Looking at the rope when looking for the snake: Conceptually mediated eye movements during spoken-word recognition. Psychonomic Bulletin & Review, 12, 453-459.

De Deyne, S., & Storms, G. (2008). Word associations: Network and semantic properties. *Behavior Research Methods*, *40*(1), 213-231.

Dell, G. S., & Chang, F. (2014). The P-chain: relating sentence production and its disorders to comprehension and acquisition. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *369*(1634), 20120394.

Duñabeitia, J. A., Avilés, A., Afonso, O., Scheepers, C., & Carreiras, M. (2009). Qualitative differences in the representation of abstract versus concrete words: Evidence from the visual-world paradigm. *Cognition*, *110*(2), 284-292.

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

Ferreira, F., Foucart, A., & Engelhardt, P. E. (2013). Language processing in the visual world: Effects of preview, visual complexity, and prediction. *Journal of Memory and Language, 69*(3), 165-182.

Fidler, F., & Loftus, G. R. (2009). Why figures with error bars should replace p-values. *Zeitschrift für Psychologie/Journal of Psychology*, *217*(1), 27-37.

Fodor, J. A. (1983). *The modularity of mind: An essay on faculty psychology*. MIT press.

Henderson, J. M., & Ferreira, F. (Eds.) (2004). *The interface of language, vision, and action: Eye movements and the visual world*. New York: Psychology Press.

Huettig, F., & Altmann, G. T. M. (2004). The online processing of ambiguous and unambiguous words in context: Evidence from head-mounted eye-tracking. In M. Carreiras, & C. Clifton (Eds.), *The on-line study of sentence comprehension: Eyetracking, ERP and beyond* (pp. 187−207). New York, NY: Psychology Press.

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

Huettig, F., & Altmann, G. T. M. (2007). Visual-shape competition during language-mediated attention is based on lexical input and not modulated by contextual appropriateness. *Visual Cognition, 15*(8), 985-1018.

Huettig, F., Rommers, J., & Meyer, A. S. (2011a). Using the visual world paradigm to study language processing: a review and critical evaluation. *Acta Psychologica, 137*(2), 151-171.

Huettig, F., & McQueen, J. M. (2007). The tug of war between phonological, semantic and shape information in language-mediated visual search. *Journal of Memory and Language, 57*(4), 460-482.

Huettig, F. (2015). Four central questions about prediction in language processing. Brain Research, 1626, 118-135.

Kamide, Y. (2008). Anticipatory processes in sentence processing. *Language and Linguistics Compass, 2*(4), 647-670.

Keuleers, E., Brysbaert, M., & New, B. (2010). SUBTLEX-NL: a new measure for Dutch word frequency based on film subtitles. *Behavior Research Methods, 42*(3), 643-650.

Knoeferle, P., & Crocker, M. W. (2006). The coordinated interplay of scene, utterance, and world knowledge: Evidence from eye tracking. *Cognitive Science, 30*(3), 481-529.

Kutas, M., DeLong, K. A., & Smith, N. J. (2011). A look around at what lies ahead: Prediction and predictability in language processing. *Predictions in the brain: Using our past to generate a future, 190-207*.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2016). *lmerTest: Tests in Linear Mixed Effects Models*. Retrieved from https://CRAN.R-project.org/package=lmerTest

Lenth, R. (2018). *Emmeans: Estimated marginal means*. Retrieved from https://CRAN.R-project.org/package=emmeans

Lupyan, G. (2012). Linguistically modulated perception and cognition: the label-feedback hypothesis. *Frontiers in Psychology*, 3, 54.

Mani, N., & Plunkett, K. (2010). In the infant's mind's ear: Evidence for implicit naming in 18-month-olds. *Psychological Science, 21*(7), 908-913.

McQueen, J. M., & Huettig, F. (2014). Interference of spoken word recognition through phonological priming from visual objects and printed words. *Attention, Perception, & Psychophysics, 76*(1), 190-200.

Masson, M. E. J., & Loftus, G. R. (2003). Using confidence intervals for graphically-based data interpretation. *Canadian Journal of Experimental Psychology/Revue Canadienne De Psychologie Experimentale, 57*(3), 203-220.

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences, 36*(4), 329-347.

Prinz, W. (1997). Perception and action planning. *European Journal of Cognitive Psychology*, *9*(2), 129-154.

Rommers, J., Meyer, A. S., Praamstra, P., & Huettig, F. (2013). The contents of predictions in sentence comprehension: Activation of the shape of objects before they are referred to. *Neuropsychologia*, *51*(3), 437-447.

R Development Core Team. (2012). R: *A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing*. Open access available at: http://cran.rproject.org

Saryazdi, R., & Chambers, C.G. (2018). Mapping language to visual referents. Does the degree of image realism matter? *Acta Psychologica, 182,* 91-99.

Saslow, M. G. (1967). Latency for saccadic eye movement. *Journal of the Optical Society of America*, *57*(8), 1030.

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*, 1632-1634.

Taylor, W. L. (1953). "Cloze procedure": A new tool for measuring readability. *Journalism Quarterly, 30,* 415–433.

Yantis, S., & Jonides, J. (1984). Abrupt visual onsets and selective attention: evidence from visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *10*(5), 601.

Yee, E., & Sedivy, J. C. (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.

Table 1: Average fixation proportions for critical objects and averaged distractors for the three display types and the two preview types, calculated for the time window starting 500 ms before target word onset and ending 200 ms after target word onset (700 ms in total).

Display type/Preview type	Pre-verb		Post-verb	
	Critical object	Averaged distractors	Critical object	Averaged distractors
Target	.25 (.3)	.18 (.26)	.32 (.31)	.11 (.21)
Semantic competitor	.19 (.27)	.19 (.27)	.31 (.29)	.12 (.25)
Visual competitor	.22 (.28)	.20 (.27)	.22 (.28)	.15 (.23)

Note: Standard deviations provided in bracket.

Table 2: Linear mixed effects model output for the analysis of eye gaze (empirical log odds) in the two preview conditions (pre-verb, post-verb) and the three display types (target-present, semantic competitor, visual competitor). Pre-verb and target-present conditions were put on the intercept.

Predictor	Coeff.	SE	t	p
Intercept	0.96	0.34	2.79	0.006
Preview_Post-verb	1.85	0.52	3.6	<.001
Display_Semantic-Competitor	-0.92	0.37	-2.47	0.014
Display_Visual-Competitor	-0.81	0.37	-2.19	0.029
Preview_Post-verb * Display_Semantic-Competitor	0.57	0.53	1.09	0.277
Preview_Post-verb * Display_Visual-Competitor	-0.79	0.53	-1.5	0.135

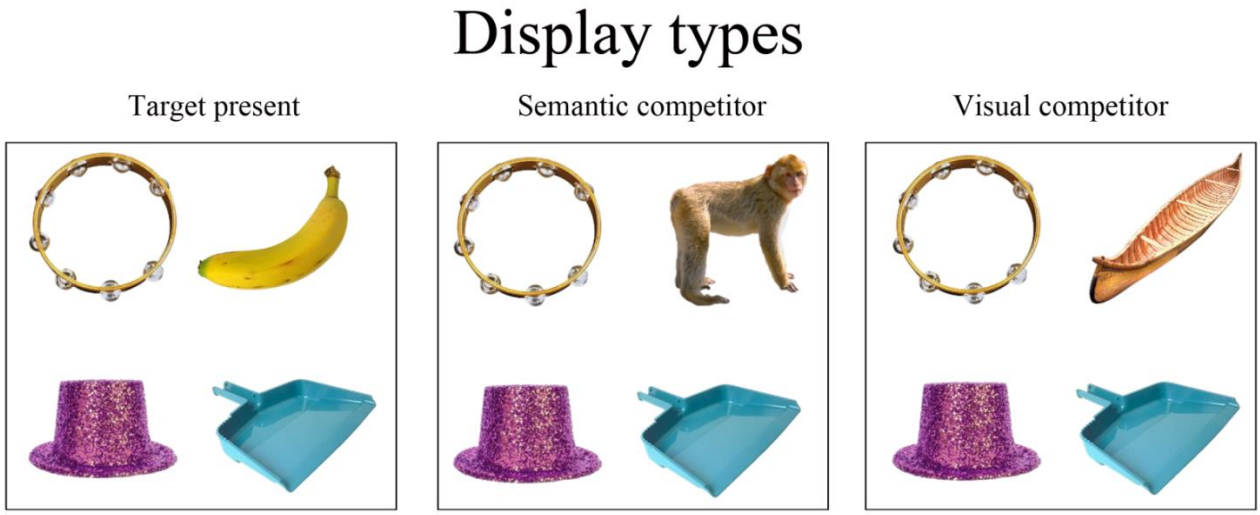
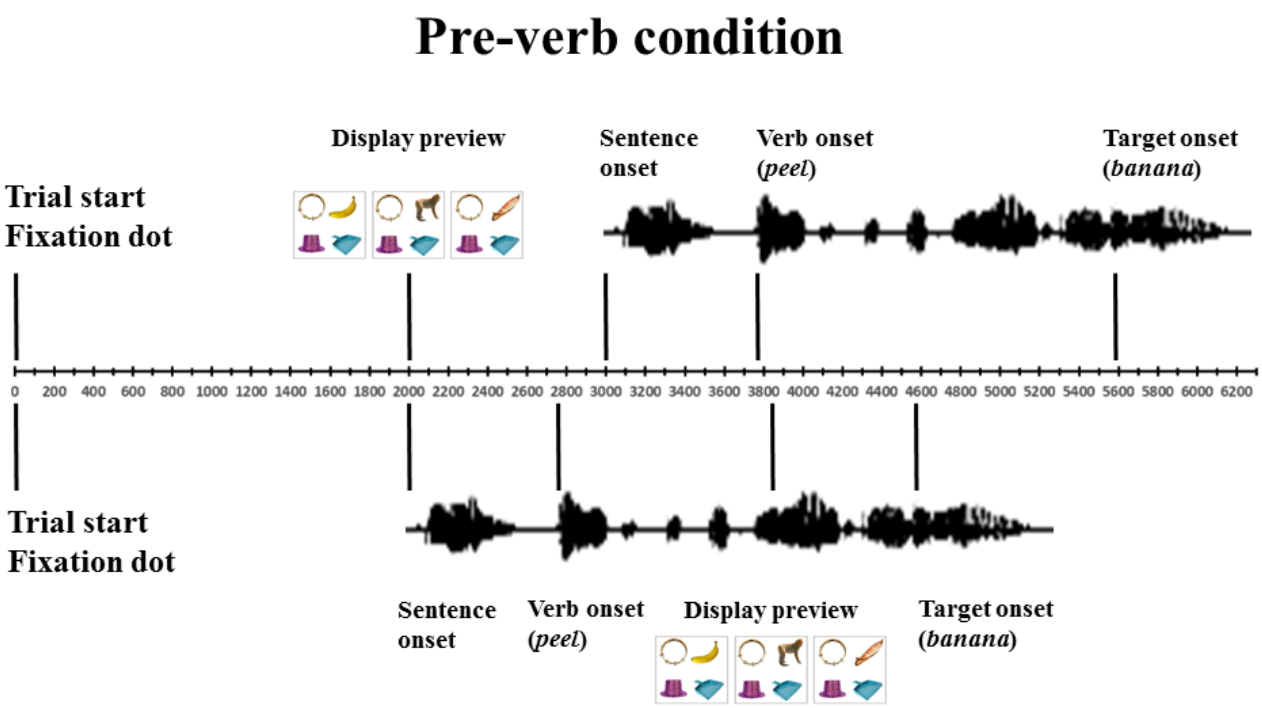


Figure 1: Examples of visual displays. While listening to the sentence, participants looked at displays in which the predictable target object was present (banana), or was absent and a semantic competitor (monkey) or a visual shape competitor (canoe) were present. In all three display types, the pictures of the hat, the tambourine and the dustpan were unrelated distractors.



556

557 Figure 2: Timeline of events in pre-verb and post-verb preview trials.

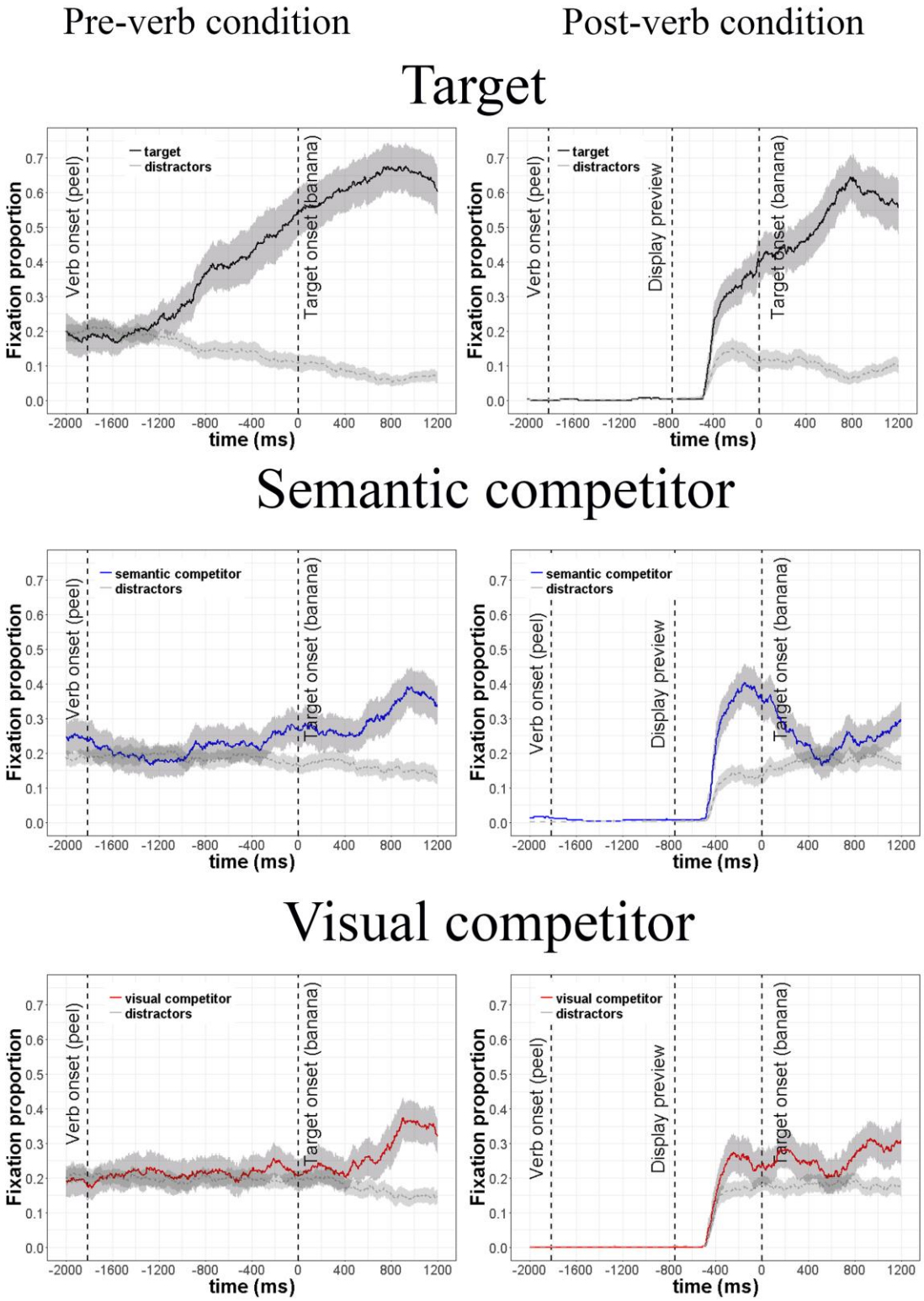


Figure 3: The graphs plot the fixation proportions for the critical objects and the averaged distractors in target-present and target-absent trials in the pre-verb and post-verb conditions. By-participant confidence intervals (95%), calculated at each sampling step, are shaded in grey.

Spoken verb	Target	Semantic comp.	Visual comp.	Distractor1	Distractor2	Distractor3
strekken (stretch)	arm (arm)	hersenen (brain)	boemerang (boomerang)	waterscooter (watercraft)	plakband (sticky tape)	koekje (cookie)
kneden (knead)	asbak (ashtray)	pijp (pipe)	jojo (yoyo)	dennenappel (pinecone)	rozen (roses)	verkeerslicht (traffic light)
pellen (peel)	banaan (banana)	aap (monkey)	kano (canoe)	tamboerijn (tambourine)	hoed (hat)	blik (dustpan)
winnen (win)	beker (cup)	vork (fork)	klos garen (bobbin)	pen (pen)	duikbril (goggles)	dynamiet (dynamite)
stapelen (stack)	blok (block)	hobbelpaard (rockinghorse)	toffee (toffee)	saxofoon (saxophone)	beer (bear)	knoop (button)
planten (plant)	boom (tree)	bijl (ax)	wc borstel (toilet brush)	magnetron (microwave)	magneet (magnet)	grammafoonspeler (gramophone player)
besturen (drive)	boot (boat)	anker (anchor)	klomp (clog)	chocolade (chocolate)	honkbal (baseball)	ventilator (fan)
branden (burn)	cd (CD)	diskette (floppydisk)	reddingsboei (buoy)	holster (holster)	meetlat (yardstick)	klamboe (mosquito net)
plukken (pick)	druif (grape)	wijnglas (wineglass)	biljartballen (billiard balls)	kettingzaag (chainsaw)	bel (bell)	megafoon (megaphone)
bakken (fry)	ei (egg)	haan (rooster)	wol (wool)	tandenborstel (toothbrush)	xylofoon (xylophone)	boog (bow)
koelen (cool)	fles (bottle)	kurk (cork)	kegel (bowling pin)	broek (pants)	kerstbal (bauble)	portemonnee (wallet)
blazen (blow/play)	fluit (flute)	harp (harp)	deegroller (rolling pin)	badeend (duck)	ton (ton)	skeeler (rollerblade)
smeden (forge)	hoefijzer (horseshoe)	zadel (saddle)	koptelefoon (headphone)	teddy beer (teddy bear)	camembert (camembert)	peultje (shell)
ontpitten (pit)	meloen (melon)	bananen (banana)	rugbybal (rugby ball)	golfclub (golf club)	slang (snake)	stoel (chair)
slijpen (sharpen)	mes (knife)	theepot (teapot)	peddel (paddle)	poederdoos (powder box)	babybedje (cot)	bokshandschoenen (boxing gloves)
drinken (drink)	milkshake (milk shake)	friet (Frenchfries)	walkietalkie (walkie talkie)	wetsuit (wet suit)	snelheidsmeter (speedometer)	pad (path)
laseren (laser)	oog (eye)	pruik (wig)	globe (globe)	broccoli (broccoli)	politieauto (police car)	stemvork (tuning fork)
piercen (pierce)	oor (ear)	voet (foot)	croissant (croissant)	schildersezel (easel)	vrachtwagen (truck)	leeuw (lion)
stemmen (tune)	piano (piano)	trompet (trumpet)	streepjescode (barcode)	riem (belt)	bureaulamp (desk lamp)	pannenkoeken (pancakes)
skimmen (skim)	pinpas (debit card)	euro (eurocoin)	envelop (envelope)	blad (sheet)	zwaan (swan)	nagelschaartje (nail scissors)
snoeien (prune)	plant (plant)	gieter (wateringcan)	feesttoeter (party horn)	wasmachine (washmachine)	controller (controller)	garnaal (shrimp)
openen (open)	raam (window)	schoorsteen (chimney)	schilderij (painting)	vishaak (fishhook)	zalmmoot (salmon fillet)	honkbalknuppel (baseball bat)
lanceren (launch)	raket (rocket)	tank (tank)	vuurtoren (lighthouse)	toilettas (toiletry)	dalmatiër (dalmatian)	vuilnisemmer (bin)
graveren (engrave)	ring (ring)	oorbellen (earrings)	donut (donut)	telraam (abacus)	prei (leek)	fluitketel (kettle)
persen (squeeze)	sinaasappel (orange)	courgette (zucchini)	golfbal (golfball)	kalf (calf)	snijplank (cutting board)	bergschoen (hiking boot)
doppen (shell)	sperzieboon (butter bean)	ui (onion)	degen (sword)	spiegel (mirror)	douchekop (showerhead)	fiets (bicycle)
knopen (tie)	stropdas (tie)	trui (sweater)	vlieger (kite)	rolstoel (wheelchair)	videoband (videotape)	notenkraker (nutcracker)
waxen (wax)	surfplank (surfboard)	badpak (swimsuit)	veer (feather)	bizon (bison)	graafmachine (excavator)	ananas (pineapple)
kleien (make pottery)	theepot (teapot)	lepel (spoon)	kandelaar (candleholder)	sportschoenen (sneakers)	bretels (braces)	ketting (chain)
spotten (spot)	vliegtuig (plane)	label (label)	kruis (cross)	worst (sausage)	muffinvorm (muffin pan)	beeldscherm (screen)

562