# Chapter 3
# The Complexity of the Visual Environment Modulates Language-Mediated Eye Gaze

**Florian Hintz and Falk Huettig**

## 3.1 Background: The Interaction of Language, Vision, and Attention During Language-Mediated Visual Search

A remarkable characteristic about human cognition is that we are able to process visual input and spoken language at the same time and have seemingly no difficulty integrating both modalities. For instance, when walking along an unknown street and simultaneously listening to someone describe the directions on the phone, we are usually able to spot all the visual landmarks mentioned by the person on the phone quickly and with ease. Within milliseconds we can locate the "statue opposite the supermarket" and eventually navigate our way through the city. Our visual search in those situations is cued by information derived from the auditory input, which is mapped on information derived from the visual surroundings. In other words, processing spoken language activates long-term linguistic and non-linguistic mental representations just as processing visual input activates associated long-term linguistic and non-linguistic representations.

F. Hintz (✉) · F. Huettig
Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
e-mail: florian.hintz@mpi.nl

F. Hintz
International Max Planck Research School for Language Sciences,
Nijmegen, The Netherlands

F. Huettig
Donders Institute for Brain, Cognition, and Behavior,
Radboud University, Nijmegen, The Netherlands

R.K. Mishra et al. (eds.), *Attention and Vision in Language Processing*, DOI 10.1007/978-81-322-2443-3_3

The general question we asked in this study is whether the nature of the visual environment has an impact on the way we process concurrent spoken language. In particular, we investigated how increased visual complexity affects the likelihood of word–object mapping at various levels of representation during language-mediated visual search. To understand how exactly language and vision interact, we must know which knowledge types are retrieved when processing both language and visual input. A useful method for the investigation of language-vision interaction is the visual world paradigm (VWP; Cooper 1974; Tanenhaus et al. 1995; see Huettig et al. 2011b, for a review). In the VWP participants hear spoken language while they look at a visual scene related to the spoken utterance. Participants' eye movements are recorded for analysis. Many recent studies have looked at the mental representations involved in word–object mapping which has led to a substantial body of literature. Allopenna et al. (1998), for example, showed that word–object mapping can occur at a phonological level of representation. The participants in that study looked at computer displays showing, for example, the pictures of a beaker (target object), a beetle (phonological onset competitor), a speaker (phonological rhyme competitor) and a carriage (unrelated distractor), while listening to the spoken instruction "Pick up the beaker". The authors observed that the participants' likelihood of fixating both the picture of the beaker and the picture of the beetle increased as they encountered the initial phonemes of the spoken target word "beaker". As the acoustic information of beaker started to mismatch with the phonological information of beetle, the likelihood of looks to the beetle decreased as the likelihood of looks to the beaker continued to rise. As the end of "beaker" acoustically unfolded, looks to the picture of a speaker started to increase. Simulations run with the TRACE model of speech perception (McClelland and Elman 1986) replicated the eye gaze pattern of the participants, consistent with the notion that the probability of fixating items within the visual display can be driven by a phonological overlap between the name of a depicted object and the target word in the auditory stimulus.

Word–object mapping can also take place at a semantic/conceptual level of representation. This has been examined in a number of studies. Huettig and Altmann (2005; see also Yee and Sedivy 2006; Yee et al. 2009; Duñabeitia et al. 2009), for instance, investigated whether semantic properties of spoken words could direct eye gaze towards objects in the visual field in the absence of any associative relationships between targets and competitors. They found that participants directed their overt visual attention towards a depicted object (e.g., trumpet) when a semantically related but not associatively related target word (e.g., "piano") acoustically unfolded, and that the likelihood of fixation was proportional to the degree of conceptual overlap (cf. Cree and McRae 2003). Similarly, Huettig et al. (2006) observed that corpus-based measures of word semantics (e.g., latent semantic analysis, Landauer and Dumais 1997) each correlated well with fixation behaviour. Based on those studies, language-mediated eye movements can be seen as a sensitive indicator of the degree of overlap between the semantic information conveyed by speech and the conceptual knowledge retrieved from the visual objects. In those experiments, phonological relationships between spoken words and visual objects were not present, hence demonstrating that semantic word–object mapping can occur in the absence of phonological mapping.

Finally, there is experimental evidence suggesting that word–object mapping occurs at a perceptual level of representation. Visual mapping, that is, increased looks to entities related for instance in visual shape, has been observed when participants were presented with the picture of a cable while listening to the spoken word "snake". The likelihood of looks to the picture of the cable increased whilst the word "snake" acoustically unfolded (Huettig and Altmann 2004, 2007; Dahan and Tanenhaus 2005). Visual mapping, likewise, was found in the absence of phonological and/or semantic mapping.

More recently, Huettig and McQueen (2007) showed that the listener's fixation behaviour during language-mediated visual search can be characterised by a *tug of war* between matches at phonological, semantic, *and* visual levels of representation. In four eye-tracking experiments, they presented participants with displays including either four visual objects (Experiments 1 and 2) or the printed word names of the same objects (Experiments 3 and 4) and concurrent spoken sentences including a critical target word which was preceded by on average seven words. The sentence preceding the critical word (e.g., "beaker") was contextually neutral (i.e., participants could not predict the target word from the sentential context). Three of the four entities in the display were related to the target word: one was related in semantics (e.g., a fork, and unrelated in phonology and visual shape), one was similar in visual shape (e.g., a bobbin, and unrelated in phonology and semantics) and the name of the third object overlapped phonologically in the first syllable with the target (e.g., beaver, and was unrelated in semantics and visual shape). In Experiments 1 and 4, participants were presented with the visual display from the beginning of the sentence but in Experiments 2 and 3, they only had a 200 ms preview of the display before the critical word acoustically unfolded. In Experiment 1, the phonological overlap between the critical spoken word and the visually presented phonological competitor object resulted in shifts in eye gaze to that object for the duration of the overlap. As the spoken target word unfolded beyond the overlapping first syllable and indicated that the competitor object was not part of the sentence, participants shifted their eye gaze to the shape and semantic competitors. When there was only 200 ms to look at the same display prior to the onset of the critical spoken word (Experiment 2), participants did not look preferentially at the phonological competitors. Instead, they made more fixations to the shape competitors and then to the semantic competitors. Huettig and McQueen (2007) interpreted the absence of an attentional bias to the phonological competitors in Experiment 2 as revealing that participants had not yet retrieved the names of the pictures before the onset of the spoken word. Hence, picture processing had not advanced to a phonological level of representation and by the time a picture name would have been retrieved, the evidence in the speech signal had already indicated that the phonological competitor was not part of the sentence. When the pictures in Huettig and McQueen (2007) were replaced with their printed word names (Experiments 3 and 4), the authors observed attentional shifts to the phonological competitors only, both with short and long previews of the display. This suggested that the likelihood of mapping between representations derived from the language input and representations derived from the visual input was contingent upon the nature of the visual stimuli (i.e. printed words vs. pictures).

Huettig and McQueen (2011) investigated this issue further and examined whether semantic and visual-shape representations are routinely retrieved from printed-word displays and used during language-mediated visual search. They used the same sentences as in the earlier study and printed word displays with no phonological competitors present. The study found evidence for semantic mapping with printed word displays (when phonological matches between the speech signal and the visual objects were not present) but not for shape mapping even though participants looked at these competitors when they were presented as pictures (Huettig and McQueen 2007, Experiments 1, 2). Huettig and McQueen argued that shape information about the objects appears not to be used in online search of printed-word displays whereas it is used with picture displays suggesting that the nature of the visual environment modulates word–object mapping.

In summary, when we are faced with spoken language and visual input at the same time, matches between representations derived from either modality can happen at phonological, semantic and visual levels of representation. The listener's fixation behaviour during language-mediated visual search seems to be determined by a tug of war between all these types of word knowledge. However, the exact level of representation at which word–object mapping takes place appears to be determined by the timing of processing in both the language and the visual processing system, the temporal unfolding of the speech signal, and by the nature of the visual environment. In the present study, we further investigated how the nature of the visual environment impacts on on-line word–object mapping.

The studies described above suggested that there are important differences in online word–object mapping between displays of visual objects and displays of printed words. Moreover, most of the research using the VWP has either used simple object displays (usually four objects, one in each corner of the screen) or its written-word equivalent. With such 'simple displays' the interpretation of fixation behaviour could only be based on the properties of the individual objects. Using complex visual scenes in those studies would have made the evaluation of the findings very difficult because the effects of scene-specific influences on fixation behaviour (e.g., scene schema knowledge) could not have been easily separated from influences of lexical effects of, e.g. semantic similarity (see for instance, the semantic influences on object identification reported by Boyce and Pollatsek 1992; De Graef 1998).

A recent study investigated word–object mapping using photographs of real world scenes (Andersson et al. 2011). Participants viewed cluttered scenes containing a large number of objects (e.g., a scene depicting a garage sale) while listening to three-sentence-passages that varied in speech rate. The authors showed that effects of language-mediated eye gaze appear to be very robust as the participants directed their visual attention to objects mentioned in the speech even under demanding conditions (e.g., a fast speech rate). What we cannot tell from those findings is at which levels of representation word–object mapping takes place in complex visual scenes because this was not part of Andersson et al.'s manipulation. As the objects mentioned in the speech signal were present in the visual scene, it is possible that matches happened at all three levels of representation (e.g., phonological, semantic, visual). However, it is also possible that word–object mapping in realistic scenes and word–object mapping in simple four-object displays are fundamentally different.

In the present series of experiments, we examined how semi-realistic visual scenes affect the likelihood of matches at phonological, semantic and visual shape levels of representation. This question has considerable real life relevance as we typically do not view objects in visually impoverished simple displays but rather in more complex surroundings. In other words, the way we experience the visual environment in our daily life is much more complex than is simulated in most experiments conducted using the VWP. In order to approximate natural language-vision interaction, we must know how exactly word–object mapping, i.e. the tug of war between phonological, semantic and visual shape information is influenced by more complex visual environments.

## 3.2 Word–Object Mapping in Complex Visual Scenes

Printed word displays have been shown to induce implicit biases during language-mediated visual search, as with such displays mappings occur mainly at phonological levels of representation. Huettig and McQueen (2011) argued that with printed-word displays, there is particularly easy access to the phonological form of words (van Orden et al. 1988; Frost 1998). Here, we tested how the increased complexity of semi-realistic scenes impacts on phonological, semantic and shape word–object mapping.

### *3.2.1 Experiment 1*

Thirty participants with normal or corrected to normal vision participated. The same visual objects were used as in Huettig and McQueen (2007).[1] We embedded those objects in semi-realistic line drawings including human-like cartoon characters with either a narrow path running through the scene or three implied walls indicating the contours of a room. Four different characters (two male and two

[1] See Huettig and McQueen (2007) for a detailed description of the materials and the results of seven norming studies. Five of the original item sets were removed from both Experiment 1 and all subsequent experiments, because they contained pictures of body parts present in the human-like characters.

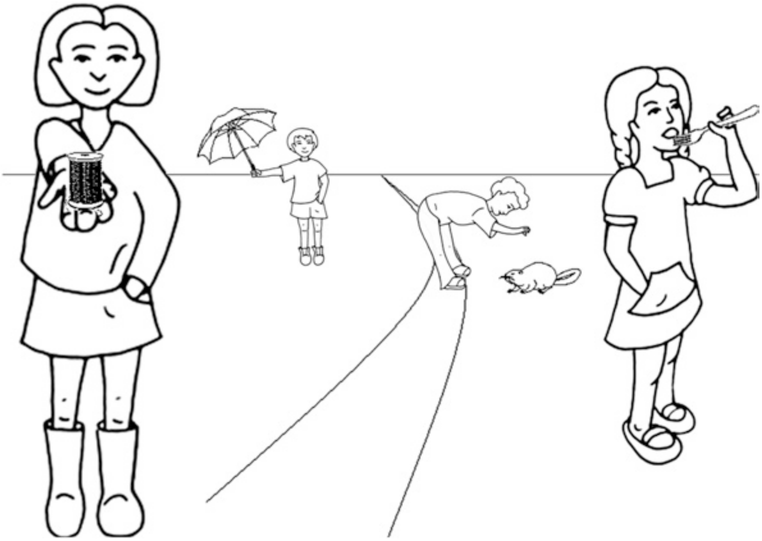

**Fig. 3.1** Example display used in Experiment 1. For the spoken word “beker”, *beaker*, the display consisted of pictures of a beaver (the phonological competitor), a bobbin (the visual-shape competitor), a fork (the semantic competitor) and an umbrella (the unrelated distractor)

female) in different postures were drawn. A random combination of four of those was present in each scene and shown to interact with the visual objects (hold, lean down to, pet, etc.). Huettig and McQueen’s (2007) stimulus materials were created to avoid any semantic relationship between the four objects. The scenes (Fig. 3.1, for an example) were hence composed such that scene schema information (i.e., contextual knowledge of objects that might be expected within a specific scene; for example, shower gel, sponge, and soap are items that one might expect to see in a bathroom scene and in particular locations, Strik and Underwood 2007) was minimised. This was done to separate effects of scene schema information from effects of the visual complexity and from effects of character–object interactions.

After a three-second preview, our participants heard single spoken target words while looking at semi-realistic scenes and were asked to indicate the presence or absence of the target object. That is, they were asked to produce “Ja” (Yes) when the target was present and “Nee” (No) when the target was absent. During filler trials, the target objects (and three unrelated distractor objects) were present, but during experimental trials they were absent and the display contained an unrelated distractor object (an umbrella, *paraplu*) and various competitor objects. For example, given the spoken target “beaker” (*beker*), the display contained a phonological (a beaver, *bever*), a shape (a bobbin, *klos*), and a semantic (a fork, *vork*) competitor.

Given the robust nature of language-mediated orienting (cf. Andersson et al. 2011), we expected that our manipulation would not result in a breakdown of word–object mapping. Compared to Huettig and McQueen (2007, Experiment 1,

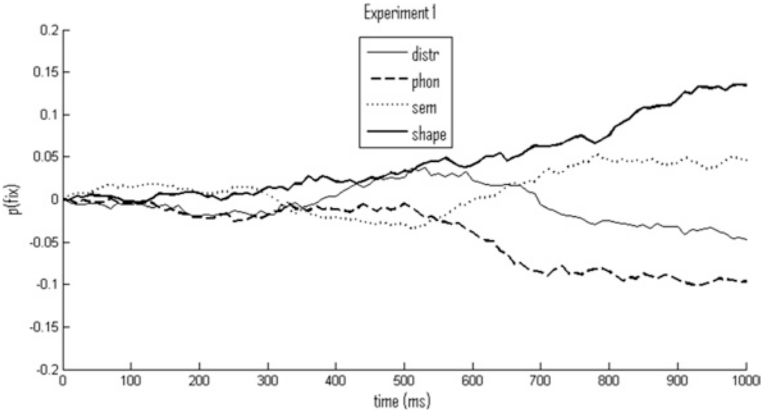

**Fig. 3.2** Time course graph showing change in fixation probabilities to phonological competitors, visual-shape competitors, semantic competitors and unrelated distractors for Experiment 1 (semi-realistic scene)

four objects in four corners of the screen), the displays in the current experiment contained more visual entities. Also, the cartoon characters were shown to interact with the visual objects possibly supplying additional semantic information, which was not present in the earlier experiment. We hypothesised that enhanced visual and semantic complexity would lead to mapping biases at semantic and visual levels at the expense of mapping at the phonological level.

Figure 3.2 shows a time course graph illustrating the change in fixation probabilities for Experiment 1,[2] for each of the four objects, for 1 s after the acoustic onset of the spoken target word. The proportion of trials with a fixation at the acoustic onset of the target word served as a baseline. Each subsequent data point reflects the proportion of trials with a fixation at that moment minus the baseline (cf. Huettig and Altmann 2005). Negative values thus reflect moves away from objects that were already fixated at the onset.

Trials on which participants had responded incorrectly were removed from the analysis.[3] For the statistical analysis, we calculated ratios between the proportion of fixations made to a particular competitor (phonological, semantic, or shape) and the sum of proportion of fixations made to the distractor object and that competitor. A ratio greater than 0.5 suggests that of all the fixations directed to a particular

[2] Prior to Experiment 1 (and Experiment 2) participants carried out an object naming task during which their eye movements were recorded. The task was independent of the subsequent main experiment and required participants to look at one object at a time presented at the centre of the computer screen and name it as fast as possible. Sixty objects which were not used in the main experiment had to be named. The task lasted around 5 min and we observed no obvious impact on participants' performance in either Experiment 1 or 2 nor did they report anecdotal effects.

[3] There was one item on which more than 50 % of the participant sample had responded incorrectly. This item was removed from further analyses, and was removed from the subsequent experiments.

competitor and the distractor, the competitor attracted more than 50 % of those fixations. Conversely, a ratio smaller than 0.5 reflects that of all the fixations directed to the competitor and the distractor, the distractor attracted more than 50 % of those fixations. Mean ratios were computed by participants and items for 100 ms time bins starting at the acoustic onset of the target word. Given the time necessary for programming and initiating an eye movement (Saslow 1967), we can assume that fixations during the 0–99 ms time window were not influenced by information from the spoken target word. Pairwise t-tests were carried out comparing the 0–99 ms bin (baseline, hereafter) to nine subsequent time bins (until 1 s after the spoken word onset). We tested, for the data in each window, whether the competitor–distractor ratio was significantly different from fixations made during the baseline. These analyses provide estimates of when competitor and distractor fixation proportions diverge (and perhaps later converge) over the time window of interest. The average duration of the spoken target words was 500 ms. The time bin analysis hence spanned the acoustic lifetime of the spoken word and additional 500 ms after the spoken word offset.

Figure 3.2 suggests a replication of the visual shape and semantic biases reported by Huettig and McQueen (2007, Experiment 1).[4] Importantly, however, there was no sign of a bias in looks to the phonological competitor as in Huettig and McQueen (2007, Experiment 1). The statistical analysis revealed that fixations to the shape competitor became significant during the time bin starting 800 ms after the spoken word onset (800–900 ms: t1(29) = −2.49, $p$ = 0.019; t2(33) = −2.29, $p$ = 0.029; 900–1000 ms: t1(29) = −3.89, $p$ = 0.001; t2(33) = −3.77, $p$ = 0.001). Fixations to the semantic competitor were significant by participants and approached statistical significance by items (t1(29) = −2.81, $p$ = 0.009; t2(33) = −1.44, $p$ = 0.159). There were no increased looks to the phonological competitors during the earlier time windows (200–300 ms: t1(29) = 1.07, $p$ = 0.296; t2(33) = 1.13, $p$ = 0.267; 300–400 ms: t1(29) = 0.88, $p$ = 0.387; t2(33) = 0.81, $p$ = 0.425). We observed, however, some evidence for inhibition of shifts in eye gaze to the phonological competitors (i.e., more looks to the unrelated distractor than to the phonological competitor) during late time bins (600–999 ms; 600–699 ms bin: t1(29) = 2.47, $p$ = 0.02); t2(34) = 1.81, $p$ = 0.079) suggesting that the phonological forms (i.e. the word names) had been retrieved.

In sum, in Experiment 1, we examined the impact of semi-realistic visual environments on the tug of war between phonological, semantic and visual shape information. We used the same materials as Huettig and McQueen (2007) but instead of presenting the visual objects in simple 2 × 2 arrays, we embedded them in semi-realistic scenes including four human-like characters, which were shown to interact with the objects. Participants showed increased fixations to visual-shape competitors. However, there was no hint of an initial bias in shifts to phonological competitors, and the bias in shifts to the semantic competitor was reduced and not statistically robust in the item analysis.

[4] Note that our main aim interest was not in the exact timing of the shifts to semantic and shape competitors. What is clear from the data (see Fig. 3.2) is that participants started to shift their eye gaze to both competitors after the target word had been heard.

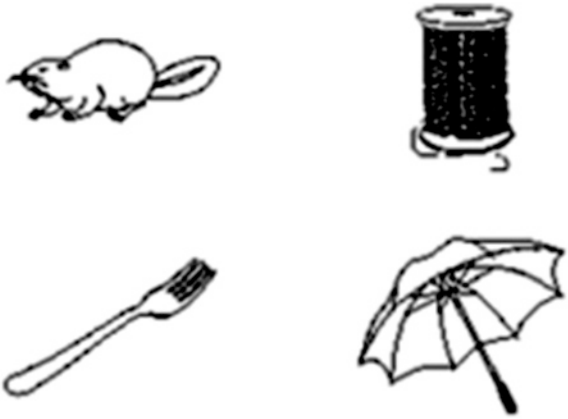

**Fig. 3.3** Example display used in Experiment 2 (cf. Huettig and McQueen 2007, Experiment 1). For the spoken word “beker”, *beaker*, the display consisted of pictures of a beaver (the phonological competitor), a bobbin (the visual-shape competitor), a fork (the semantic competitor), and an umbrella (the unrelated distractor)

Experiment 2 was conducted to rule out an alternative explanation namely that the differences between the present results and those of Huettig and McQueen (2007, Experiment 1) were due to differences in the tasks used. Huettig and McQueen (2007) asked participants to simply look at the displays while listening to the spoken language. In the present study, we asked participants to pursue an active search task, i.e. to indicate (by saying yes or no) whether the visual object referred to by the spoken target word was present in the display or not. Hence, we cannot rule out that the observed differences in the results between the two studies were due to task differences. In Experiment 2, we presented participants with the same simple object arrays as used in Huettig and McQueen (2007, Experiment 1, cf. Fig. 3.3, for an example) but instructed them to carry out the same task as in the present Experiment 1. If the differences in results are due to task differences, then the data of Experiment 2 should be similar to the results of the present Experiment 1. If on the other hand the difference in the nature of the visual environment is crucial, Experiment 2 should replicate the data pattern of Huettig and McQueen (2007, Experiment 1).

### 3.2.2 *Experiment 2*

Thirty subjects who had normal or corrected to normal vision and had not participated in Experiment 1 were tested in Experiment 2. The results revealed that task differences are unlikely to account for the differences between the present Experiment 1 and Experiment 1 of Huettig and McQueen (2007). Figure 3.4 and the statistical analyses show that with simple four object displays, using an active task, participants’ fixations to the phonological competitor objects (300–400 ms: $t1(29) = -2.3$,

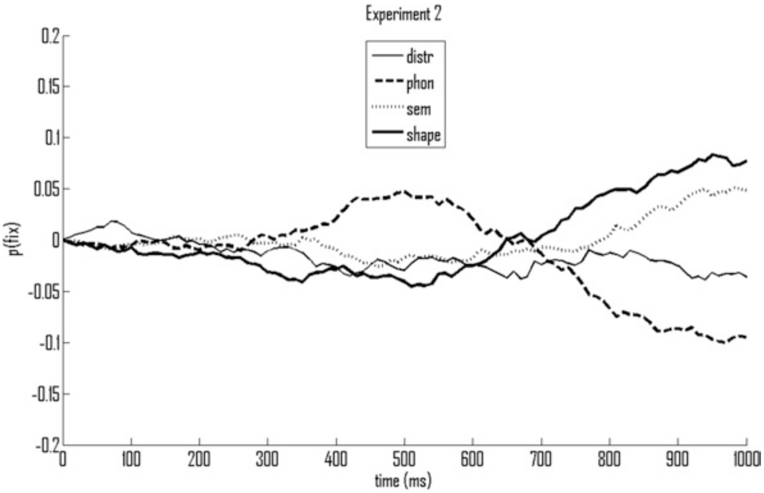

**Fig. 3.4** Time course graph showing change in fixation probabilities to phonological competitors, visual-shape competitors, semantic competitors and unrelated distractors for Experiment 2 (simple object displays)

$p = 0.029$; t2(33) $= -2.96$, $p = 0.006$) preceded those to shape (e.g., 900–999 ms: t1(29) $= -2.91$, $p = 0.007$; t2(33) $= -2.15$, $p = 0.039$) and semantic competitors (e.g., 900–999 ms: t1(29) $= -2.81$, $p = 0.009$; t2(33) $= -1.44$, $p = 0.159$). The results of Experiment 2 were therefore a close replication of Huettig and McQueen (2007, Experiment 1). Hence, we can rule out that the observed differences between the present Experiment 1 and Experiment 2 were due to an active task being used instead of a passive look-and-listen task. More likely, these differences can be attributed to the varying nature of the visual environment in those experiments.

The findings are intriguing, indicating substantial differences in word–object mapping between semi-realistic scenes and 2 × 2 object arrays. The data pattern in Experiment 1 suggests that participants show an increased preference for visual mapping with the increased visual complexity of the semi-realistic scenes. Why might participants show a preference of visual mapping during language-vision interactions involving complex visual scenes?

One possibility is that in the more complex display, the objects were less salient (i.e., took more time to find) than in the simpler displays. This could have had two implications: First, this could delay the retrieval of the objects' phonological representations such that by the time the onset of the spoken words occurred, picture processing had not cascaded to levels at which phonological forms are retrieved (cf. Huettig and McQueen, Experiment 2). We believe that to be unlikely. Participants had sufficient preview of the visual scenes (3 s) before the spoken words were heard. Perhaps more importantly, our data show some evidence for phonological inhibition during the 600–999 ms time window. This suggests that the participants had retrieved the phonological forms of the objects.

Second, one might argue that the lack of visual salience of the objects in the semi-realistic scene affected the time-course of the mapping process of language-derived and vision-derived representations. That is, when the target word (e.g., “beaker”) acoustically unfolded, participants were not able to locate the phonological competitors quickly. Note that such an account predicts that all types of competitors should be equally affected. However, this was not the case. We found a strong and robust bias for mapping at the visual level of representation and a clear (albeit reduced) tendency for semantic mapping.

What then are the mechanisms modulating word–object mapping in semi-realistic scenes? One could argue that our semi-realistic scenes did not substantially increase the ‘semantic content’ in the displays as the character–object couplings were rather arbitrary and all four character–object pairs did not belong to a semantically coherent scene, thus limiting the extent of semantic mapping. However, as indicated above, the employment of visual displays, where scene schema information is present, is generally difficult as lexical effects cannot be differentiated from effects of scene schema knowledge. In the semi-realistic scenes, visual complexity however was substantially increased as compared to the 2 × 2 displays. If visual complexity resulted in the visual bias then we should expect a replication of the pattern of Experiment 1, even if the characters are replaced with meaningless shapes.

### 3.2.3 *Experiment 3*

In Experiment 3, we therefore, removed the character–object interactions by replacing the human-like drawings with unnamable, meaningless black shapes. Those shapes have been used in earlier studies on statistical learning of higher order temporal structures (cf. Fiser and Aslin 2002, see Fig. 3.5, for an example display used in the current study). That way, the additional four visual entities cannot be interpreted as interacting with the objects, yet we kept the scene visually complex. In all other respects, the set-up was identical to Experiment 1. Thirty subjects who had not participated in either Experiment 1 or 2 were tested.

The fixation graph in Fig. 3.6 and the statistical analysis[5] revealed that while there were robust biases to shape (700–799 ms: t1(29) = −3.43, $p$ = 0.002; t2(32) = −2.13, $p$ = 0.041) and semantic (800–899 ms: t1(29) = −2.92, $p$ = 0.007; t2(32) = −2.01, $p$ = 0.053) competitors, there was not a tendency for looks to the phonological competitors. As in Experiment 1, we found some evidence for phonological inhibition (t1(29) = 2.74, $p$ = 0.01; t2(32) = 1.39, $p$ = 0.175). These results are consistent with our predictions that increased visual complexity induces a bias of word–object mapping at the visual level of representation.

[5] Due to an error, one experimental item had to be removed from the analysis.

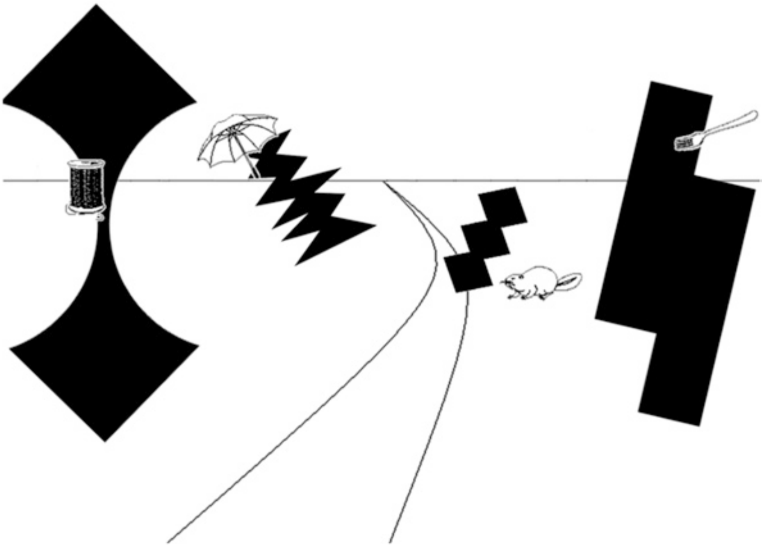

**Fig. 3.5** Example display used in Experiment 3. For the spoken word “beker”, *beaker*, the display consisted of pictures of a beaver (the phonological competitor), a bobbin (the visual-shape competitor), a fork (the semantic competitor) and an umbrella (the unrelated distractor). The cartoon characters were replaced with meaningless *black shapes*

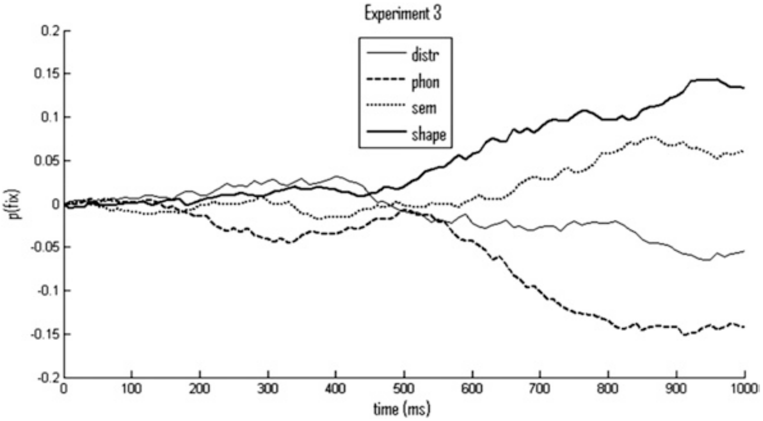

**Fig. 3.6** Time-course graph showing change in fixation probabilities to phonological competitors, visual-shape competitors, semantic competitors and unrelated distractors in Experiment 3 (cartoon characters replaced with meaningless shapes)

## 3.3 General Discussion

The present findings provide further evidence that during language-mediated visual search with picture displays, there is a tug of war between multiple types of mental representations (e.g., phonological, semantic, visual shape, cf. Huettig and McQueen 2007). Our main aim in the present study was to assess the influence of more complex visual environments on this tug of war. To this end, we showed identical visual objects either in more complex visual environments including four human-like cartoon characters or four meaningless black shapes, or as simple four object arrays. Participants heard single spoken target words while looking at the different displays and were asked to indicate the presence or absence of the target objects. We hypothesised that the more complex visual information and the semantic information intrinsic to the semi-realistic scenes would induce a mode of processing yielding matches at the visual and semantic level. We assumed that this mode of processing would lead to a modulation of mapping behaviour at the phonological level, in other words, to a reduced likelihood of phonological mapping.

In Experiment 1, we observed an attentional bias in looks to the visual-shape competitors and a tendency for a bias for the semantic competitors, but there were no shifts in eye gaze towards phonological competitors when the objects were embedded in semi-realistic scenes. When the objects were presented in simple four object displays (Experiments 2), however, we replicated the clear early attentional bias to phonological competitors found in earlier research (Huettig and McQueen 2007, Experiment 1). This showed that task differences (active vs. look-and-listen task) could not account for the absence of shifts in attention to phonological competitors with semi-realistic scenes. Crucially, we observed fixation behaviour very similar (biases for visual-shape and semantic mapping) to that in the present Experiment 1, when the human-like cartoon characters in the earlier experiment were replaced with meaningless black shapes (Experiment 3).

This suggests that the pattern of results was not driven by the objects being presented in interaction with the human-like cartoon characters but can most likely be attributed to the increased visual complexity in the scene.

### *3.3.1 Why Do More Complex Visual Environments Reduce the Likelihood of Word–Object Mapping at the Phonological Level of Representation?*

While increasing the complexity of the visual scene, we re-arranged the objects' regular distribution over the display. Usually, those are arranged in a square and the distances between all objects are the same. In the semi-realistic scenes we used, this was not the case (compare Figs. 3.1 and 3.3). One could argue that this might have affected the mapping process as saccades might have been longer or shorter as compared to 'regular-arranged' object displays. But if a regular arrangement of objects in a symmetrical square was crucial for the observed eye gaze behaviour, all types of competitors should be affected. However, for the visual-shape and semantic competitors, we observed similar biases as in the experiments with simple object displays.

We conjecture that the mode of processing towards mapping of visual-shape and semantic features of objects in the present experiments with more complex visual scenes was induced by the increased amount of visual information present in the visual scene. That is, increased visual processing led participants to a mode where matches at visual (and semantic) levels are preferred over matches at the phonological level. The character–object formations (Fig. 3.1) and the black shape–object formations (Fig. 3.5) are visually more complex than the same objects being presented in isolation. That is, with more visual information present in the displays, visual processing was enhanced, shifting word–object mapping preferences.

Importantly, Experiment 3 showed that the mere presence of additional visual entities is sufficient to induce this mode of processing and was more important in modulating word–object mapping than the semantic information inherent to the character–object couplings. In fact, those might have even hindered the extent of semantic mapping as they were not contributing to a semantically coherent scene. Võ and Wolfe (2013) recently showed that viewing semantically altered scenes elicits electrophysiological responses similar to when semantically implausible sentences are comprehended (e.g., N400 deflections). We suggest that semantically fully coherent visual scenes (e.g., a kitchen scene) may induce an enhanced bias towards word–object mapping at a semantic level of representation. Future work could usefully explore this hypothesis.

### *3.3.2 Inhibition of Phonological Word–Object Mapping*

In Experiment 1 and Experiment 3 we found some evidence for inhibition in looks to the phonological competitors, i.e. fewer looks to the competitor than to the distractor during late time windows. This pattern might reveal general insights into eye gaze behaviour during language-mediated visual search in more complex visual environments. This is interesting because to our knowledge effects of inhibition during language-mediated search have previously not been demonstrated (but see McQueen and Huettig 2014, for evidence from three cross-modal priming experiments for interference of spoken word recognition through phonological priming from visual objects).

The current results strongly suggest that language-mediated eye movements are (at least partially) under substantial control processes and that a complete account of language-mediated eye gaze will have to include inhibitory mechanisms. We suggest that processing in the current experiments was contingent upon attentional control over how processing is distributed across different levels of representation (cf. Stolz and Besner 1998). Indeed substantial amounts of cognitive control during language-mediated visual search have recently been predicted by a working memory model of language-vision interactions (Huettig et al. 2011a). The authors proposed that working memory plays a central role during language-mediated eye movements, because it grounds linguistic, cognitive, and perceptual processing in space and time by providing short-term connections between objects (cf. Knoeferle and Crocker 2007; Spivey et al. 2004). If language-attention interactions are mediated by working memory, such interactions are likely to be subject to a substantial amount of cognitive control. Han and Kim (2009) showed recently in a (non-language) visual search task that although working memory appears to bias visual selection towards matching stimuli, participants exerted some control over which items are ignored in the search display especially when search was slow. As language-mediated visual search tends to be slower than a standard visual search task, it is likely to be under increased cognitive control. Future research could usefully examine the nature of the inhibition effects observed in the present study and explore underlying mechanisms and conditions in which they occur.

### *3.3.3 Conclusion*

Given our results, one may ask to what extent word–object mapping at the phonological level of representation occurs during real world language–vision interactions. A strong conclusion from our data would be that in complex visual surroundings word–object mapping at a phonological level of representation is the exception rather than the rule and limited to situations with very simple visual environments. However, such a conclusion may be premature. Our results indicate that the dynamics of the representational level at which online word–object mapping occurs is determined by, among other things, the complexity of the visual environment. There are many other factors (e.g., cascaded processing in the spoken word and picture recognition systems; the temporal unfolding of the spoken language, the particular task goals, etc.) also co-determining this mapping behaviour. The present findings do not rule out that there are situations in which mapping at a phonological representational level is particularly potent even in complex visual environments. Another mediating factor for instance appears to be literacy skills. Huettig et al. (2011c) found robust evidence for word–object mapping at the phonological level in high literates. In low literates (who had no reading or other cognitive impairments), however, they observed that word–object mapping (with four object displays) takes place primarily at the semantic level.

In sum, word–object mapping is contingent upon the nature of the visual environment. More complex visual environments induce visual modes of processing during language-mediated visual search. The data suggest further that word–object mapping is under substantial cognitive control.

**Acknowledgments** We thank Anna Gastel for drawing the semi-realistic scenes, Neil Bardhan for providing the artificial shapes and Johanne Tromp for assistance in running the experiments.

## References

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

Andersson, R., Ferreira, F., & Henderson, J. M. (2011). I see what you're saying: The integration of complex speech and scenes during language comprehension. *Acta Psychologica, 137*(2), 208–216.

Boyce, S. J., & Pollatsek, A. (1992). Identification of objects in scenes: The role of scene background in object naming. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 18*(3), 531–543.

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

Cree, G. S., & McRae, K. (2003). Analyzing the factors underlying the structure and computation of the meaning of chipmunk, cherry, chisel, cheese, and cello (and many other such concrete nouns). *Journal of Experimental Psychology: General, 132*(2), 163–201.

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 and Review, 12*(3), 453–459.

De Graef, P. (1998). Prefixational object perception in scenes: Objects popping out of schemas. In G. Underwood (Ed.), *Eye guidance in reading and scene perception* (pp. 313–336). Oxford, UK: Elsevier.

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.

Fiser, J., & Aslin, R. N. (2002). Statistical learning of higher-order temporal structure from visual shape sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*(3), 458–467.

Frost, R. (1998). Toward a strong phonological theory of visual word recognition: True issues and false trails. *Psychological Bulletin, 123*(1), 71–99.

Han, S. W., & Kim, M. S. (2009). Do the contents of working memory capture attention? Yes, but cognitive control matters. *Journal of Experimental Psychology: Human Perception and Performance, 35*(5), 1292–1302.

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: 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*(1), 23–32.

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., & 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., & McQueen, J. M. (2011). The nature of the visual environment induces implicit biases during language-mediated visual search. *Memory and Cognition, 39*(6), 1068–1084.

Huettig, F., Olivers, C. N. L., & Hartsuiker, R. J. (2011a). Looking, language, and memory: Bridging research from the visual world and visual search paradigms. *Acta Psychologica, 137*(2), 138–150.

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.

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

Huettig, F., Singh, N., & Mishra, R. (2011c). Language-mediated visual orienting behavior in low and high literates. *Frontiers in Psychology, 2*, 285.

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

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.

McClelland, J. L., & Elman, J. L. (1986). The trace model of speech perception. *Cognitive Psychology, 18*(1), 1–86.

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

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

Spivey, M. J., Richardson, D. C., & Fitneva, S. A. (2004). Thinking outside the brain: Spatial indices to visual and linguistic information. In J. Henderson & F. Ferreira (Eds.), *The interface of language, vision, and action: Eye movements and the visual world* (pp. 161–190). San Diego: CA: Academic Press.

Stolz, J. A., & Besner, D. (1998). Levels of representation in visual word recognition: A dissociation between morphological and semantic processing. *Journal of Experimental Psychology: Human Perception and Performance, 24*(6), 1642.

Strik, J. A., & Underwood, G. (2007). Low-level visual saliency does not predict change detection in natural scenes. *Journal of Vision, 7*(10), 1–10.

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*(5217), 1632–1634.

Van Orden, G. C., Johnston, J. C., & Hale, B. L. (1988). Word identification in reading proceeds from spelling to sound to meaning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*(3), 371.

Võ, M. L.-H., & Wolfe, J. M. (2013). Differential electrophysiological signatures of semantic and syntactic scene processing. *Psychological Science, 24*(9), 1816–1823.

Yee, E., Overton, E., & Thompson-Schill, S. L. (2009). Looking for meaning: Eye movements are sensitive to overlapping semantic features, not association. *Psychonomic Bulletin and Review, 16*(5), 869–874.

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.