Chapter 1

Introduction
INTRODUCTION

There are multiple ways in which we gather information about our environment. For humans both vision and speech are important sources of information. Interestingly, we can easily integrate the two modalities. For example, when hearing a word, people can rapidly direct their gaze to the correct referent in the visual environment, even when the environment contains many other objects. Somehow we are able to direct our attention to, and select the correct visual object amongst competing information on the basis of language. But how do we this? How do we integrate vision and language in order to select objects in the environment?

One view, and one which is dominant in the visual attention literature, is that people select a relevant object amongst competing other objects on the basis of a visual image (also called attentional template, Desimone & Duncan, 1995; Wolfe, 2007). So when people hear for example the word “cup”, they will imagine what a cup looks like, i.e. activate its visual representation. This visual template then biases selection towards a matching object in the visual environment. Thus selection is then considered to be visually driven.

However, visual representations are not the only type of representation that can be activated. Words, of course have meaning, and the same is true for visual objects. For example, when looking at a cup, additional knowledge can be retrieved, such as that it is meant to hold fluid, that you can drink from it and that it can normally be found in a kitchen. This is referred to as semantic information. As it is possible to activate semantic representations from visual and linguistic input, selection could thus in principle also be semantically driven.

Although there is some empirical evidence in the visual attention literature that semantics can influence selection (which will be discussed extensively in the section “Visual attention literature: the visual search task” below), most important models of visual attention do not incorporate semantics as a guiding property (e.g., Itti & Koch, 2001; Wolfe, 2007). In contrast, in a different field of psychology, psycholinguistics, it is well recognized that visual objects and words can activate multiple levels of representation, and that orienting can thus take place on the basis of all these levels (see the section “Psycholinguistics: the visual world paradigm” below).

The main question of this thesis is when are objects selected on the basis of visual representations (like shape, texture and color) and when are semantic representations (like function and meaning) more relevant for
selection? To answer this question, knowledge from visual attention studies and psycholinguistic research will be combined. Unfortunately, the communication between these disciplines has been scarce, even though these fields could learn much from each other (Huettig, Olivers, & Hartsuiker, 2011). For example, visual attention researchers have mainly focused on the visual aspects of selection, but have not extensively acknowledged the influence of other levels of representation like semantics. In this literature there is a long tradition of systematically investigating selection mechanisms, like attentional capture and delayed disengagement, and exploring the influence of other factors, for example cognitive control. In psycholinguistics, on the other hand, researchers have not investigated the specific mechanisms behind selection or looked at factors like cognitive control, but instead have extensively examined how other levels of representation like semantics can drive selection and how these influences might change over time (i.e., they considered dynamic rather than static measures).

Two paradigms are especially important for the current topic, namely the visual world paradigm from the psycholinguistic field and the visual search task from the visual attention literature. This chapter will therefore start with a short review of the relevant visual world studies followed by an overview of the relevant visual search studies. Finally, several general theoretical frameworks will be discussed. From these frameworks several predictions can be made which formed the basis of this thesis.

Psycholinguistics: The visual world paradigm

The interaction between visual and linguistic input has been extensively investigated in psycholinguistics with the visual world paradigm (Cooper, 1974; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995; see Huettig, Rommers, & Meyer, 2011, for an extensive review). Here participants see semi-realistic scenes or arrays of visual objects or printed words. After some time a spoken utterance is presented. Participants are either instructed to manipulate the objects (so called direct action tasks, e.g., Allopenna, Magnuson, & Tanenhaus, 1998) or to passively look and listen (so called passive viewing task, e.g., Altmann & Kamide, 1999; Huettig & Altmann, 2005; Huettig & McQueen, 2007). Eye movements are continuously recorded while the speech unfolds. Researchers are interested which objects are being fixated over time as
fixations are considered to reflect a closely time-locked measure of ongoing cognitive processes. Specifically, the dependent variable proportion fixation time (“P(fix)” in short) is used. This is the proportion of time people spend fixating a particular object within a certain time period. When P(fix) is plotted over multiple time bins, it clearly shows how fixation preferences change over time. Traditionally, the visual world paradigm has been used as a tool to explore linguistic processes (Huettig, Rommers, et al., 2011), like for example speech perception (e.g., Allopenna et al., 1998; Salverda, Dahan, & McQueen, 2003), disfluency (e.g., Arnold, Fagnano, & Tanenhaus, 2003), and bilingual word recognition (e.g., Spivey & Marian, 1999; Weber & Cutler, 2004). However, the concurrent presentation of visual and linguistic information and the possibility of exploring the dynamics through eye movements make the visual world paradigm also useful for studying the influence of language on visual orienting.

Three linking hypotheses

In the visual world literature, researchers have acknowledged that vision and language can interact at multiple levels of representation. Some have proposed that this mainly happens on the basis of semantic representations. This semantic mapping hypothesis was supported by an early finding of Cooper (1974). This study found that participants, while listening to short narratives, did not only look at the correct referent, but also fixated objects that were semantically related to a critical word. So after hearing the word “Africa” participants were more likely to look at pictures of a snake, a zebra and a lion than at unrelated objects. Over the past few years this result has been replicated multiple times (e.g., Dunabeitia, Aviles, Afonso, Scheepers, & Carreiras, 2009; Huettig & Altmann, 2005; Yee, Overton, & Thompson-Schill, 2009; Yee & Sedivy, 2006). Additionally, it has also been found that several corpus-based measures of word semantics (McRae feature norms, Cree & McRae, 2003; latent semantic analysis, Landauer & Dumais, 1997; contextual similarity, McDonald, 2000) predict fixation behavior (Huettig, Quinlan, McDonald, & Altmann, 2006).

At the same time, other studies showed that objects that are visually related to a critical word also have a higher chance of being fixated than unrelated objects. Thus after hearing the word “snake” people are more prone to fixate objects that looked like a snake, for example a cable (e.g., Dahan &
This has led to a visual mapping hypothesis that states that language is translated into a visual representation and that the two modalities interact at a visual level of representation. Finally, besides a semantic and a visual mapping hypothesis, researchers have also found evidence supporting a phonological mapping hypothesis. According to this hypothesis both the linguistic and visual stimuli activate a phonological representation and the two modalities then interact at a phonological level. This idea is supported by the findings that people fixate more often objects that share the first phoneme (“beaker-beetle”) or rhyme with the critical word (“beaker-speaker”) (e.g., Alloppena et al., 1998; McQueen & Huettig, 2012; McQueen & Viebahn, 2007) and that objects’ names are rapidly activated when people search for a specific item (e.g., Meyer, Belke, Telling, & Humphreys, 2007; Gorges, Oppermann, Jescheniak, & Schriefers, 2013; Walenchok, Hout, & Goldinger, 2016).

Importantly, Huettig and McQueen (2007) proposed that these three types of mapping might not be mutually exclusive: At which level of representation vision and language interact is flexible. In their study three of the four objects in the display were related to the speech input: One at the semantic level, one at the visual level, and one at the phonological level. The corresponding sentence could for example contain the word “beker” (beaker), while the visual display contained a picture of a beaver (“bever”, a phonological related object), a bobbin (a visually related object), a fork (semantically related object) and an umbrella (unrelated object). Pictures were presented at sentence onset (on average after 6.85 other words), so participants could typically preview the pictures for a long period of time before the critical word was presented. Participants initially fixated the phonologically related objects more often than the other objects in the display (phonological bias), but over time this phonological bias disappeared, and participants spent more time fixating visually and semantically similar objects than unrelated objects (respectively a visual and semantic bias). In a second experiment, the pictures were presented only 200 ms before critical word onset rather than at sentence onset. Participants thus did not have as much time to preview the pictures. Now no phonological bias was observed anymore, but there were again visual and semantic biases. Support for the phonological mapping hypothesis was thus only found when the pictures were previewed for a long time, i.e. when people had enough time to retrieve the
names of the pictures in the display. Besides the timing, Huettig and McQueen (2007) found that orienting biases were also affected by the nature of the display. When the display contained words instead of pictures, a reliable orienting bias was found towards phonologically related words, but there were no orienting biases towards visually or semantically related words (see Huettig & McQueen, 2011 for an argument that these biases are present but arise later). Overall, the results of Huettig and McQueen (2007) suggest that multiple factors - like timing and the nature of the input - determine at which level of representation language and vision meet - a conclusion that this thesis further subscribes. The present thesis however will focus on how the meaning conveyed by language affects selection of visual objects and therefore the phonological component will not be examined.

As a side note, the studies discussed so far only examined orienting biases that occurred after critical word onset. However, another line of research has found that specific word structures before the critical word can already direct participants’ attention to certain objects (Kamide, Altmann, & Haywood, 2003; Altmann & Kamide, 1999, 2007, 2009). For example, Altmann and Kamide (1999) presented sentences were the verb was predictive for an upcoming noun. In two different versions of a sentence a boy could either move or eat the cake. The visual display contained a picture of a cake besides a picture of a boy and several non-edible objects. Participants significantly fixated the cake earlier when the sentence contained the verb “eat” than when it contained the verb “move”. In addition, researchers have also found anticipatory eye movements towards objects that were only visually or only semantically related to a critical word (Hintz, Meyer, & Huettig, submitted; Rommers, Meyer, & Huettig, 2015; Rommers, Meyer, Praamstra, & Huettig, 2013). In the current thesis single words were used instead of sentences. Anticipatory eye movements will thus not be discussed further.

**Blank screen paradigm: Orienting on the basis of memory**

In the standard version of the visual world paradigm, the visual input is always present during the spoken utterance. However, in an important variant, the blank screen paradigm, the visual stimuli are removed prior to word onset. Multiple studies have shown that when the objects are visually no longer present and are thus only available in memory, people still orient towards locations previously occupied by the referred word, even when this was
unnecessary for the task (e.g., Altmann, 2004; Dell'Acqua, Sessa, Toffanin, Luria, & Jolicoeur, 2010; Hoover & Richardson, 2008; Johansson & Johansson, 2014; Laeng & Teodorescu, 2002; Richardson & Kirkham, 2004; Richardson & Spivey, 2000; Spivey & Geng, 2001; Theeuwes, Kramer, & Irwin, 2011). So after hearing the word “rabbit” people will look at the location where earlier a picture of a rabbit was presented. A way to interpret these “looks at nothing” is by assuming that subjects have formed episodic memory traces where visual object identities are bound to their respective locations. When the target object is then being mentioned, its original location will also be retrieved, which in turn triggers an eye movement towards this location (Ferreira, Apel, & Henderson, 2008; Richardson, Altmann, Spivey, & Hoover, 2009).

Although studies have shown that people fixate locations previously occupied by the referred to object, it has been so far unclear whether the same eye movement behavior would occur with objects that are only semantically or only visually related to the referred objects. This was therefore investigated in Chapter 5 of this thesis.

**Visual attention literature: The visual search task**

In the visual attention literature researchers are interested in how objects are selected from a rich visual environment that contains competing information. Linguistic input is then a way to inform the participant about the relevant object, rather than a topic of investigation in itself. According to the biased competition model (Desimone & Duncan, 1995), observers set up a perceptual template (or attentional set) in working memory, which is used to bias attention towards matching items in the visual environment. A common way to investigate selection processes is with the help of the visual search task (Wolfe, 1998 for a review). In this type of task participants search within a visual display for a pre-specified target object amongst other, distractor objects. In visual search tasks, the verbal or visual target instruction is usually presented before the onset of the search display, and often even before the start of a block or session. Participants then get the explicit task to respond as quickly and accurately as possible to specific features of the target, its location or its presence (Huettig, Olivers, et al., 2011). Often the total number of objects in the display (set size) is manipulated and search times are then considered as a function of set size, with the slope of this function then providing a measure of
INTRODUCTION

efficiency. Besides response times, researchers have also used eye movement measures (e.g., proportion of first fixations, saccade latency time and fixation duration), EEG measures (e.g., ERP components such as the N2pc and Pd) and fMRI. Many studies have used alphanumeric or simple stimuli (like Gabor patches or shapes), while some have used line drawings or photos of real-life objects or scenes.

Importantly, whereas in psycholinguistics researchers believe that multiple types of representation can serve object selection (resulting in the various mapping hypotheses), the dominant view in the visual attention literature is that selection happens on the basis of visual representations (for reviews, see Theeuwes, 2010; Wolfe & Horowitz, 2004). Most models of visual search therefore do not incorporate any other form of semantic guidance (e.g., Itti & Koch, 2001; Wolfe, 2007). However, such models deny visual search studies that indicated that templates can also operate on a semantic level. Next, several visual search studies will be discussed that looked more into the question whether the search template can be semantic in nature.

Categorical search
A semantic contribution to attentional guidance has been shown by studies that examined whether people can distinguish between visually similar objects purely on the basis of categorical information. For example, in Jonides and Gleitman (1972) participants had to indicate whether an ambiguous “O” target was present or absent. Half of the participants were told that the target was the letter “o”, whereas the other half heard that it was a zero (digit). People either searched for the target amongst letters or digits. This created within-category search (e.g., “look for the letter amongst other letters”) or between-category search (e.g., “find the letter amongst digits”). For within-category search, researchers found steep search slopes indicating serial search. Surprisingly, shallow search slopes were observed for between-category search suggesting parallel guidance towards the target. Search was thus more efficient when the ambiguous target was presented amongst items from another category (but see for non-replications Duncan, 1983; White, 1977). Additionally, Lupyan (2008) found that ps were faster detected than bs when presented amongst Bs, even though bs and ps are visually similar, and share the same number of features with the B. Apparently, the fact that the B distractors belonged to the same category as the b target caused extra
interference. Additional conceptual information can thus alter the search for visually similar targets.

**Visual versus verbal target instructions**

One line of research has focused on the question whether search on the basis of visual instructions (pictorial cues) is as efficient as with verbal instructions (word cues, which has been considered as a more semantic way of providing instructions). Here it has been found that verbal instructions can guide attention, but never as efficiently as pictorial instructions (e.g., Castelhano, Pollatsek, & Cave, 2008; Knapp & Abrams, 2012; Malcolm & Henderson, 2009; Maxfield & Zelinsky, 2012; Meyers & Rhoades, 1978; Schmidt & Zelinsky, 2009; Smith, Redford, Gent, & Washburn, 2005; Vickery, King, & Jiang, 2005; Wilschut, Theeuwes, & Olivers, 2014; Wolfe, Horowitz, Kenner, Hyle, & Vasan, 2004; Yang & Zelinsky, 2009).

However, the fact that verbal instructions can guide visual attention does not necessarily mean that this happens on the basis of a linguistic representation. It could still be the case that both pictorial and verbal instructions are turned into a visual representation. A picture then contains more detailed information about the target than a word describing the same thing, therefore resulting in a more specific representation. This would then result in a benefit for visual over verbal instructions.

Indeed studies have found stronger attentional guidance with verbal instruction that contained more detailed information. In Schmidt and Zelinsky (2009) the verbal instruction either described the target object more precisely (“boots”) or more abstract (“footwear”). In some conditions a specific color of the object was added (“brown footwear” or “brown boots”). The results showed that more detailed verbal information increased the amount of attentional guidance. Maxfield and Zelinsky (2012) also found that search guidance was better when the instruction was presented on a subordinate level (“chocolate ice cream”) than on a basic (“ice cream”) or superordinate (“dessert”) level. Also when the depicted object in the instruction had a slightly different orientation than the actual target in the display search times were also increased (Vickery et al., 2005; see for other ways of decreasing target precision and the same conclusions, Hout & Goldinger, 2015), although search remained always faster than search on the basis of a verbal cue. Moreover, search has been found to be more efficient with typical than with atypical
objects (Castelhano et al., 2008; Maxfield, Stalder, & Zelinsky, 2014; Robbins & Hout, 2015). A closer resemblance between the actual target and the instruction thus leads to an increased search performance.

In sum, although verbal instructions can guide attention, it remains unclear from these studies whether this guidance is based on the semantic information that it conveys, or whether the verbal instruction is simply translated into a visual representation, and guidance then occurs on the basis of this visual image.

**Guidance by visual versus verbal working memory representations**

Researchers have also concentrated on the question whether an irrelevant representation in working memory can guide attention. This is normally investigated with the help of a dual task. Here people first memorize a certain object which is only relevant for a later memory task (e.g., the color red). During the retention period participants perform a search task where they are instructed to look for another object (e.g., a diamond amongst circles). On some trials the memory item will then be presented in the search display as a distractor (e.g., a red circle). If the items in working memory guide attention – despite their irrelevance to the search, then these distractors will capture attention and search will be slowed down.

Studies have indeed shown that irrelevant items in working memory can capture attention when people have to memorize visual objects (e.g., Olivers, Meijer, & Theeuwes, 2006; Olivers, 2009; Soto, Heinke, Humphreys, & Blanco, 2005), but whether this can also happen on the basis of a more verbal or more semantic representation has been debated. For example, Soto and Humphreys (2007) found that both an actual depiction of a red square as well as the words “red square” slowed down search when one of the distractors in the search display was red. Recently, Sun, Shen, Shaw, Cant, and Ferber (2015) have shown that this might also work with semantically related words like rose for a red distractor. On the other hand, Olivers et al. (2006) provided evidence that guidance depends on the representational nature of the memory. They did not alter the input of the working memory content, but rather changed the memory task so that it forced people to have either a more visual or more semantic representation. Participants always had to remember a specific color, but the memory task differed, such that observers had to make a forced choice between different color categories (i.e., red amongst blue and yellow) or
between different hues of the same color (i.e., all objects were red). Whereas in the first task people could rely on a more categorical representation, a more visual representation was needed in the latter. The results showed that the matching distractor in the visual search task only captured attention when people were forced to have a more visual representation in memory (see also Dombrowe, Olivers, & Donk, 2010). Given the contradictory findings, it thus remains a question whether irrelevant verbal or semantic representations in working memory can guide attention.

**Including semantically related distractors in the search display**

If the search template would include semantic information, one would expect that objects that are semantically related to the target instruction would draw attention and would thus slow down search. Indeed, Moores, Laiti, and Chelazzi (2003) showed that when people were looking for, for example, a motorbike, their attention was also drawn to a picture of a helmet. This semantically related object hindered the decision whether the target was absent and attracted more first fixations (see also Belke, Humphreys, Watson, Meyer, & Telling, 2008; Meyer et al., 2007). Moreover, semantically related distractors have also been shown to affect the N2pc, a relatively early EEG marker of selection (Telling, Kumar, Meyer, & Humphreys, 2010). The N2pc was enhanced when the target and the semantically related picture were presented on the same side of the display compared to a condition where the distractor was absent. When the target and the distractors were presented at opposite sites, the N2pc was reduced.

In the studies discussed above, the distractor was semantically related to the target. However, an interesting question is whether attention would also be captured by a distractor that is semantically related to an irrelevant object in working memory. Calleja and Rich (2013) instructed participants to memorize pictures for a later recognition task. During the retention period, participants searched for another object. One of the distractors in the display could be exactly the same object people had to memorize or could be semantically related. The results showed that visual search was slower when the distractor completely matched the object in working memory, but not when it was only semantically related.

Together these studies suggest that semantically related objects attract attention when they match the target instruction (i.e., when they are...
related to an item that is task relevant), but not when there is a match with an item in working memory (i.e., an item that is task irrelevant). In Chapter 4 this prediction was tested more directly.

**Semantics within scenes**

In another field, researchers have looked at semantic influences within scenes. One topic that has been extensively investigated is how the general setting of the scene (i.e., scene gist) affects what we select in our environment, rather than the meaning of standalone objects (see for reviews, Henderson & Hollingworth, 1999; Oliva, 2005). Other studies have concentrated on the influence of the semantics of individual objects within a scene. For example, Hwang, Wang, and Pomplun (2011) found that the eyes preferred objects in the scene that were semantically related to each other (e.g., people tend to look at the fork when they just fixated a knife) or semantically related to a target object (e.g., people tended to look at a fork when they were looking for a knife), comparable to what others have found with standalone objects (Moores et al., 2003; Belke et al., 2008).

The influence of semantics in scenes will not be extensively discussed within this thesis (see for an elaborate review, Wu, Wick, & Pomplun, 2014) as all our studies used displays of standalone objects. However, Chapter 6 will dive deeper into the discussion, raised in the scene perception and search literature, whether it is possible to process the semantics of objects when these are presented outside foveal vision (>2°), more specifically in peripheral vision (>5°). This can be investigated by using scenes that contain objects that are semantically incongruent with the current scene (i.e., octopus in a farmyard). To notice the inconsistency people must process the scene gist as well as the semantics of the object. Some studies have found that these incongruent objects immediately attract attention, indicating that objects’ semantics presented in extrafoveal vision can be processed (e.g., Becker, Pashler, & Lubin, 2007; Bonitz & Gordon, 2008; LaPointe & Milliken, 2016; Loftus & Mackworth, 1978; Underwood et al., 2007; Underwood, Templeman, Lamming, & Foulsham, 2008). However, this finding and associated claims have not been undisputed (e.g., De Graef, Christiaens, & D'Ydewalle, 1990; Gareze & Findlay, 2007; Henderson, Weeks, & Hollingworth, 1999; Võ & Henderson, 2009, 2011). However, the data presented in this thesis also suggest that extrafoveal semantic processing is possible under some circumstances, for
example when standalone objects are being used instead of scenes. The data of Chapter 3 and Chapter 4 was re-analyzed in Chapter 6 to see whether this claim could be made more explicit.

**General theoretical frameworks**

Two general theoretical frameworks of visual-linguistic interactions have been proposed. One of them is the *cascaded activation model of visual-linguistic interactions* (Huettig & McQueen, 2007) which states that people activate different levels of representation when they hear a word or see an object. The activation of these representations is cascaded, which means that there is a clear order in activation, but one process can start before the other ends. This model assumes that when people see an object, the visual representation will be activated before the associated semantic and phonological representations. For hearing a word the information flow is the other way around. Here people are confronted with the phonological structure first (i.e., the sounds), and therefore a phonological representation will be activated before the associated semantic and visual representations.

However, from the cascaded activation model it is not clear whether these representations are activated automatically or which cognitive systems are involved. Another general theoretical framework, the working memory hypothesis (Huettig, Olivers, et al., 2011) has been more explicit about this. According to this framework working memory plays a crucial role in visual-linguistic interactions. Long term memory is also considered to be important as stable information about the objects is being stored here. However, the environment in daily life, as well in most visual search and visual world studies, has a strong temporary component: The exact location of objects varies over time. Working memory might then be involved in dynamically binding the stable information from long term memory to the more temporary information about the environment. This hypothesis relates to ideas in the blank screen literature (Ferreira et al., 2008; Richardson et al., 2009), the idea that visual objects are “instantiated” (i.e., long-term representations (types) are bound to a specific location marked by an object file, token, or index, Kahneman, Treisman, & Gibbs, 1992; Kanwisher, 1987; Pylyshyn, 2001; see also Hoover & Richardson, 2008), the episodic buffer in the working memory model of Baddeley (2000) and with ‘embodied’ cognition views of working memory.
INTRODUCTION

(Postle, 2006; Spivey, Richardson, & Fitneva, 2004; Wilson, 2002).

In sum, both the cascaded activation model and the working memory model assume that the activation of different representations is cascaded. But additionally the working memory model is more explicit about the cognitive systems that might be involved. Both frameworks lead to a number of testable predictions which formed the basis of this thesis and which will be discussed in more detail next.

Predictions derived from the frameworks

Time. One prediction directly derived from these frameworks is that over time vision and language will match at different levels of representation. Therefore, the main dependent variable within this thesis is the proportion fixation time (“P(fix)”) rather than more standard visual search eye movements measures, like first fixation, fixation duration and saccade latency. With P(fix) it is possible to see clearly whether fixation preferences change over time.

A related prediction is that the timing of the stimuli presentation determines at which level of representation vision and language meet. In the visual search task the (auditory) instruction is normally presented before visual stimuli onset. There is thus ample time to activate all the associated representations of the word before the visual stimuli appear. According to the cascaded activation model, the visual stimuli will first activate visual representations (as they are visual in nature) and only later to associated semantic representations. Because of this cascaded activation, visually related objects will thus be fixated earlier in time than objects that are only semantically related to the word (i.e., a visual bias would precede a semantic bias). However, in visual world studies the order of stimuli presentation is often the other way around: here visual stimuli normally precede the word. In visual world studies there is thus enough time to activate both visual and semantic representations of the visual objects. When the word is then presented, a visual and a semantic representation will be activated around the same time. So in such a condition vision and language can interact at both visual and semantic levels of representation. Here visual and semantic biases will thus be more similar in their temporal dynamics.

These predictions were tested in Chapter 3 of this thesis by contrasting a condition that resembled a classical visual search study (i.e., auditory target
instruction preceded the visual stimuli) with a condition that was more like a standard visual world study (i.e. the target instruction was presented after visual stimuli onset). Additionally, it was examined to what extent the semantic bias arose independently of the visual bias.

**Cognitive control versus automaticity.** When working memory plays a crucial role in visual-linguistic interactions, as proposed in the working memory hypothesis of Huettig, Olivers, et al. (2011), then it is expected that these interactions would be subject to cognitive control, as this is one of the key features of working memory. However, in the past it has been argued that these interactions are fully automatic (Mishra, Olivers, & Huettig, 2013) as they tend to be fast, unconscious and overlearned (cf. Logan, 1988; Moors & De Houwer, 2006). Also, within the visual world paradigm, the use of a passive or an active task did not seem to yield different results (Hintz & Huettig, 2015, Experiment 2).

A passive and an active task might differ in their instructions, but this does not necessarily mean that participants do something different in both tasks. It is not harmful to perform an active task in the passive viewing studies. Therefore participants could in principle ignore the instructions and still do an active task instead. A better way to investigate whether the activation of different levels of representation are partly under cognitive control is to actively instruct participants to perform another task. In Chapter 4 this was done by manipulating the task relevance of the word. Here participants heard a spoken word before the search display appeared. This word had to be remembered for a subsequent memory task. In one condition this word was also relevant for the search as it defined the target, whereas in the other condition it was irrelevant as participants were instructed to look for another object. Note that task performance in the irrelevant condition would suffer when participants would not follow the instructions. If cognitive control is involved, than it is expected that the associated visual and semantic representations will be activated less or differently in the irrelevant condition than in the relevant condition. After all, visual and semantic representations are not needed in the irrelevant condition. However when these representations are activated automatically, the same results are expected for the irrelevant and relevant condition.
INTRODUCTION

In the absence of visual stimuli. The working memory hypothesis claims that the associated representations will be tied to a visual object and its location when given sufficient time. Activation of one source of information will thus lead to activation of all temporarily related information (see also Ferreira et al., 2008; Richardson et al., 2009). Evidence in favor of this idea comes from the blank screen paradigm (or memory search), where visual stimuli are removed prior to word onset. Here it has been found that after hearing a word, participants fixate locations that were previously occupied by the target. However, with the target there is a match at all levels of representation. In Chapter 5 it was explored whether participants would also fixate locations previously occupied by an object that was only semantically or only visually related, so whether a match on only one level of representation would be sufficient to also activate the location information of the object and thus trigger an eye movement.

A new stimulus set

None of the experiments in this thesis could have been conducted without an extensive and properly controlled stimulus set. Different stimulus sets have been published in the literature. However, many of these sets contain only semantic relationships (e.g., Belke et al., 2008; Huettig & Altmann, 2005; Kovalenko, Chaumon, & Busch, 2012; Moores et al., 2003; Telling et al., 2010; Yee & Sedivy, 2006), or only visual relationships (e.g., Dahan & Tanenhaus, 2005; Huettig & Altmann, 2007; Rommers et al., 2013; Rommers et al., 2015) between words and pictures. With these sets it is thus impossible to compare the influence of different representations against each other. The few sets that do include both semantic and visual relationships tend to be small in size (e.g., Huettig & McQueen, 2007). Researchers have moved around this problem by either using between-participant designs or by repeating specific items during the experiment. Both solutions are disadvantageous as testing between participants leads to a decrease in power and efficiency, whereas repeating stimuli could result in unwanted effects of learning and familiarity. Another drawback of earlier used stimulus sets is that they often contain (line-) drawings (e.g., Belke et al., 2008; Huettig & McQueen, 2007; Telling et al., 2010). Compared to photos of real life objects, drawings are less detailed and rich in terms of the number of perceptual dimensions, reducing ecological
validity. A final drawback is that many published sets are often not sufficiently controlled. If normed at all, stimulus set included only norms for either visual or linguistic aspects. This is because traditionally visual search researchers have mainly been interested in visual aspects, whereas in psycholinguistics researchers have focused on linguistic factors. When investigating the interactions between modalities, it is better to control on both factors.

In Chapter 2 a new stimulus set is introduced that obviates the above mentioned drawbacks. It contains photos of real life objects that are normed and controlled on both linguistic and visual aspects. Moreover, each word is paired to an object that is semantically (but not visually) related and one object that is visually (but not semantically) related. Additionally, each word was also paired to two unrelated objects. The set contains 100 of such word-picture pairs, thus including far more stimuli than the sets used in earlier studies. For the later chapters, this set was extended with another 20 trials. This made it possible to test different conditions within rather than between participants. To introduce a target absent-present task, another 120 trials were included in which the object referred to by the word was actually present.

Summary and conclusion

This general introduction reviewed different studies from the visual world and visual search literature focusing on the question whether semantics can influence visual selection. In both fields, there seems, at least, some evidence that semantics can guide selection. However, the factors that influence whether semantics guides selection have not been explored systematically. In this thesis several factors were explored, like the timing of the stimuli presentation (Chapter 3), task relevance (Chapter 4) and the absence of the visual stimuli (Chapter 5). Overall, this thesis will show that the question whether visual orienting is driven by semantics is better rephrased as to when do semantics drive visual orienting.