Building Event Meanings from Linguistic and Visual Representations: Evidence from Eye Movements

Julia Di Nardo

A Thesis

In the Department

of

Psychology

Presented in Partial Fulfillment of the Requirements

For the Degree of Doctor of Psychology at

Concordia University

Montreal, Quebec, Canada

December 2010

CONCORDIA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

This is to certify that the thesis prepared

By: **Julia Di Nardo**

Entitled: **Building Event Meanings from Linguistic and Visual Representations: Evidence from Eye Movements**

and submitted in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY (Psychology)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

_____ Chair
Dr. H. Muchall

_____ External Examiner
Dr. D. Titone

_____ External to Program
Dr. D. Isac

_____ Examiner
Dr. N. Segalowitz

_____ Examiner
Dr. M. von Grunau

_____ Thesis Supervisor
Dr. R. de Almeida

Approved by _____
Dr. A. Chapman, Graduate Program Director

December 3, 2010

_____
Dr. B. Lewis, Dean
Faculty of Arts and Science

ABSTRACT

Building Event Meanings from Linguistic and Visual Representations: Evidence from Eye Movements

Julia Di Nardo, Ph.D.
Concordia University, 2010

This paper reports three studies that were conducted to explore how the meaning of events is constructed through the use of spoken language comprehension and dynamic scene information. These studies served as an extension of prior work conducted by the present author with the overarching aim of modifying aspects of the methodology to modulate the salience of the linguistic context. Participants eye movements were recorded as they watched short movie clips of everyday events and listened to sentences related to those events. Specifically, we were interested in measuring how quickly the target object in each scene (the grammatical referent of the main verb in each sentence) would be fixated after the verb was uttered, and in some version of the movies, after the agent initiated movement toward the target object (synchronized with the verb onset, known as the disambiguating point). The first experiment replicated a previous study (Experiment 2 in Di Nardo, 2005) with the aim of increasing the visual angle of the scenes, thus reducing attentional shifts without corresponding eye movements. Experiment 2 investigated whether the absence of the linguistic context would alter the pattern of eye movements across the scene. Experiment 3 tested the hypothesis that introducing a more semantically salient initial clause to the sentences would lead to faster eye movements to the target object. Results showed that increasing the visual angle did not serve to reduce saccade onset times (SOTs), although the presence of spoken language did shorten SOTs even in the context of agent motion toward the target object.

Finally, introducing semantically restrictive initial clauses produced a mixed pattern of results; SOTs were reduced in the absence of agent motion when the verb was not semantically restrictive, although SOTs were longer when the agent moved toward the target object and the verb was not semantically restrictive. Results are discussed within the framework of the Coordinated Interplay Account (Crocker, Knoeferle, & Mayberry, 2009).

Acknowledgements

I would first like to thank my dissertation supervisor, Dr. Roberto de Almeida, for his support and intellectual input to the development of this work. For creating the software needed to run and analyse data from the eye-tracker, I would like to thank Nabil Khoury. My gratitude also extends to the laboratory's research assistants, for their help in recruiting and running participants. In particular, I am indebted to Alexandra Marquis for her assistance in collecting a large part of the data for Experiment 3. Thanks also to Linnea Stockall for inspiring the second of the experiments presented here.

I am also grateful for the financial support I have received, in the form of a graduate fellowship from FQRSC, as well various sources of supplementary grants and bursaries from Concordia University and grants held by my supervisor. This has allowed me to pursue the highest of educational pursuits with firm financial footing.

I would also like to thank my parents, who always knew I would someday be a "doctor." Their lessons have taught me that working hard and striving for excellence will get you anywhere you want to go. I am confident that these lessons will continue to serve me well in my future endeavours.

Last, but certainly not least, I wish to thank my husband, Bobby. Your love and support are what helped me cross that finish line. You held the vision for me when the going got tough, and put your priorities on hold so I could accomplish my goals. For that, I dedicate this volume to you.

Table of Contents

List of Figures

List of Tables

Building Event Meanings from Linguistic and Visual Representations: Evidence from Eye Movement Behaviour

Spoken language comprehension in naturalistic contexts, what Crocker, Knoeferle, and Mayberry (2010) have termed situated sentence processing, involves the complex interaction of the visual, linguistic, and both long-term and working memory systems. The nature of this complex interaction has been subject to much debate in the last decade or so, with early studies (e.g., Altmann & Kamide, 1999) often framing the question within Fodor's (1983) modularity theory. At issue was whether the visual and linguistic systems operate independently at the early stages of sentence comprehension, with the outputs of those systems being integrated in a later centralized cognitive system (modular theory), or whether there is evidence for early, incremental and interactive processing between the two systems (interactive theory). More recent studies, however, have begun to frame their findings in terms of recursive models that account for this complex interaction between the multiple systems (e.g., the coordinated interplay account; see Crocker et al., 2010) rather than in the dichotomous terms stipulated by modularity theory.

The so-called visual world paradigm has served as the methodological testing ground for the debate surrounding the interaction of language and vision (Henderson & Ferreira, 2004). This paradigm involves recording participants' eye movements as they attend to and foveate various objects (often references of linguistic tokens) while listening to unfolding utterances related to those objects. The central assumption of this paradigm is that the pattern of eye movements reflects the online processing of situated language comprehension, which is mediated by the visual-attentional system, and

therefore reflects the precise nature of the interaction between the visual and linguistic systems.

Perhaps a more meaningful issue in the study of situated language processing is how the grounding of language in the visual environment allows for the construction of event meanings through the integration of visual and linguistic representations. While discourse is not always related to the immediate visual context, the acquisition of language in young infants, for instance, is highly dependent on reference to objects in the environment (e.g., Dunham, Dunham, & Curwin, 1993; Harris, Jones, Brookes, & Grant, 1986; Pruden, Hirsh-Pasek, Golinkoff, & Hennon, 2007). As language skills develop, discourse related to the immediate environment not only makes reference to particular individual entities (e.g., naming objects), but also the relationship between those entities. These relationships often involve events that occur both across time and space and are lexicalized as verbs. Therefore, the study of situated language processing, with an emphasis on verbs in particular, can reveal how the visual and linguistic systems interact during event comprehension.

How the representations of these events are constructed is the central motivation of the research presented here. It involves the unique study of event-related language processing embedded in dynamic visual scenes, which can inform our current understanding of the cognitive architecture responsible for the interaction of the memory, visual and linguistic processing systems. We present three experiments that explore the pattern of eye movements produced during the viewing of these scenes, and the processing of spoken sentences related to those scenes. These experiments manipulate the salience of the visual and linguistic contexts with the aim of determining how the

linguistic stream affects visual search patterns. The first experiment replicates previous work by the present author (Di Nardo, 2005) to determine whether an increase in the visual angle of the scenes would increase the visual-attentional system's sensitivity to verb-thematic constraints—i.e., attention to potential real-world referents of grammatical objects. The second compared the pattern of eye movements when the language context was either present or absent to determine whether the linguistic stream has any effect on guiding visual fixation. The third experiment attempted to strengthen the initial semantic context of the utterance by "priming" the object to be named in the subsequent main clause. Before reporting these experiments in detail, a discussion of key visual-world studies which have made use of the technique employed here will be elaborated below.

Visual World Studies

Since the groundbreaking work of Yarbus (1967) and Cooper (1974), a number of studies employing the visual-world methodology have gone on to explore the control of visual attention in situated language processing. In a series of studies by Tanenhaus and his colleagues, static scene perception and sentence comprehension were shown to be incrementally interactive (Chambers, Tanenhaus, Eberhard, Filip, & Carlson, 2002; Sedivy, Tanenhaus, Chambers, & Carlson, 1999; Spivey, Tanenhaus, Eberhard, & Sedivy, 2002; Spivey, Tyler, Eberhard, & Tanenhaus, 2001; Tanenhaus, Magnuson, Dahan, & Chambers, 2000; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). These studies mainly focused on the recording of eye movements as participants were asked to manipulate real objects arrayed on a table. In one study, Tanenhaus and colleagues (Tanenhaus et al., 1995; see also Eberhard, Spivey-Knowlton, Sedivy, & Tanenhaus, 1995, and Spivey et al., 2002) contrasted sentences that were either

syntactically ambiguous or unambiguous with respect to the visual referents of the objects named in the sentences. For example, eye movement behaviour in response to the two instructions, *Put the apple on the towel in the box* (syntactically ambiguous) and *Put the apple that's on the towel in the box* (unambiguous) was compared when there was an apple already on a towel, an apple on a napkin, and an empty towel in the visual display. They found that eye movements were launched towards the object referents approximately 250 ms after the nouns were uttered. In addition, the time course of eye movements was closely locked to the utterance of words that would aid in resolving ambiguity, in order to establish reference as quickly as possible. Tanenhaus et al. (1995) take these results as support for the incremental interaction of the visual context with language interpretation at its early stages (in this case, syntactic parsing) used to help to establish reference in real-world displays.

In a similar series of studies, using the same methodology (tracking eye movements as participants are asked to manipulate objects in a real-world display), Sedivy and her colleagues (Sedivy et al., 1999) attempted to extend these findings to the resolution of *semantic* ambiguities. In order to examine how and when visual context mediates the resolution of semantic ambiguities, participants were asked to touch the visual referent of a named object present within an array of objects as their eye movements were recorded. They manipulated the number and order of adjectives preceding the nouns, as well as the objects in the array; for example, participants were asked to *Touch the yellow comb* when there were two yellow objects in the array, one of which was the comb. They demonstrated that the interpretation of adjectives was incremental, as indicated by the pattern of participants' eye movements; namely, that

ambiguities were resolved as the sentence unfolded. As with Tanenhaus and colleagues' studies, they take these results to indicate that contextual and linguistic information interact at the very early stages of sentence comprehension.

While the studies reviewed thus far used the visual world paradigm to investigate ambiguities in syntax and semantics, the utterances employed were usually command-based. These create a cognitive set in which the participant is actively seeking out objects in the real world display to meet the demands of the task set, rather than passively viewing them. This likely influences both the time course and scan path of eye movements. Other, more recent visual world studies have used sentences that are declarative rather than command-based, while still referring to the entities and events depicted by the scenes. Specifically, recent studies have begun to examine explicitly the role of verbs in guiding visual attention to the referents of their grammatical complements, with the aim of further investigating the interactivity debate (e.g., Altmann & Kamide, 1999; Boland, 2005; Knoerferle & Crocker, 2007). Without any specific task demand, the information encoded by verbs can be allowed to influence gaze patterns. The role of verbs as linguistic markers of events, the semantic and syntactic information they encode, and how the study of verbs can inform our understanding of situated language processing in the visual world paradigm is reviewed next.

The Role of Verbs in the Visual World Paradigm

Verbs are semantically and syntactically complex categories. Not only do they make reference to actions and states, but they also encode information about verb arguments (grammatical subject and direct object of the verb, as well as thematic roles such agent, patient and theme) and possible adjuncts. Embedded within a visual context,

verbs therefore act as linguistic markers for the relationship between various entities in the scene. Upon their utterance, they can also serve to direct visual attention to the referents of their objects within the scene.

Work conducted by Altmann and his colleagues used the visual-world paradigm to study the role of semantic information encoded by verbs. Specifically, they tested the hypothesis that verb-specific semantic information can guide visual attention towards the object referent of the verb's direct object, *before* the noun itself is uttered. Altmann and Kamide (1999), in particular, presented participants with line drawings containing a person surrounded by several objects. For example, in one scene, there is a boy sitting on the floor surrounded by an array of objects, including a toy train and a cake. Participants' eye movements were recorded as they scanned these scenes and listened to sentences related to those scenes—for example, *The boy will eat/move the cake.* The sentences differed with respect to the verb used: one was more semantically restrictive (*eat*) than the other (*move*). Crucially, the direct object of the semantically restrictive verb only had one possible visual referent in the scene; that is, in the example given here, the cake was the only edible object in the array. In contrast, the verb *move* could have referred to any of the objects within the scene.

Altmann and Kamide (1999) were interested in two aspects of participants' eye movement behaviour: first, what proportion of saccades were launched towards the visual referent of direct object of the verb (cake) before the onset of the noun, and second, at which point in time participants first launched a saccade towards the target object relative to the onset of the verb. They found that participants fixated the target object in 90% of the trials, and that the first saccade to the target object was launched prior to the onset of

the noun in 38% of the semantically non-restrictive trials and in 54% of the semantically restrictive trials. However, there was no significant effect of verb type. With regards to the saccade onset time, they found that the first saccade after the verb occurred 127 ms after the onset of the noun in the semantically nonrestrictive condition, and 85 ms before the onset of the noun in the semantically restrictive condition. Here, the effect of verb type was significant. These data suggest that verb-specific information does direct eye movements towards objects that are semantically consistent with the selectional restrictions of the verbs.

However, task demands may have affected the speed and pattern of eye movements, in that the judgment of whether the sentence applied to the scene or not required at the end of every trial may have induced anticipatory eye movements. Thus, a second experiment was carried out without such a judgment task (Altmann & Kamide, 1999, Experiment 2). Here, the findings were similar: the target object was fixated in 93% of the trials, and the first saccades towards the target object were initiated before the noun in 18% of the non-restrictive trials and in 32% of the restrictive trials (this difference was statistically significant). In this experiment, the saccade onset time was 536 ms in the non-restrictive condition and 591 ms in the restrictive condition (this difference was also statistically significant). Notably, saccade onset times were much longer in this experiment, as might have been expected due to the lack of the decision task. Nevertheless, the data still point to the ability of verb-specific information to guide eye movements to the visual referents of complement noun phrases.

Altmann and Kamide (1999) speculate that these results suggest that the language processing system can predict, based on objects activated by the scene, possible fillers of

a verb's patient role (the entity upon which the verb acts) prior to the utterance of the noun phrase. This is consistent with McRae, Ferretti, and Amyote's (1997) and Dowty's (1991) thematic role assignment theory, where thematic roles are presumed to contain world knowledge about typical agents (the entity that carries out the action denoted by the verb) and patients. Importantly, they hypothesize that when role concepts are activated, candidate noun fillers are actively sought out by the visual-attentional system and evaluated for their compatibility in fulfilling the semantic restrictions of the verbs.

A series of experiments conducted by Kamide, Altmann, and Haywood (2003) examined whether verb arguments other than *themes* (patients, or direct objects), such as *goals* (indirect objects), could guide eye movements in the same way. Using line drawings accompanied by spoken sentences, they found that more eye movements were directed towards objects that were consistent with the *goals* of the verbs uttered in the sentences than those that were not. Again, this suggests that verbs contain semantic/syntactic information that can be used to predict, or at least constrain, possible role fillers. In addition, it also supports the idea that scene knowledge is integrated with ongoing linguistic processes such as verb-semantic role assignment.

While the studies conducted by Altmann and his colleagues have focused on the semantic properties of verbs, Boland (2005) focused on the syntactic properties of verb complements using the visual world paradigm. In particular, she contrasted verb arguments with verb adjuncts (complements that are not required by the verb but rather serve to elaborate the action being expressed). In addition, unlike Altmann and colleagues (Altmann & Kamide, 1999; Kamide et al., 2003), her studies employed pictures of objects rather than drawings. In one experiment, she presented participants

with sentences using a dative verb (e.g., *suggest*) with a *Recipient* argument that was either typical (e.g., *teenager*, in …*the mother suggested [the newspaper] to her teenager*) or atypical (e.g., …*to her toddler*). These dative constructions were contrasted with sentences whose nouns were adjuncts of the verb (e.g., *Instruments* as in…*the farmer beat [the donkey] vigorously with a stick/hat*…). The results showed that there was a greater proportion of fixations to the visual referents of these nouns when they were part of a verb's argument structure (e.g., the datives) than when they belonged to adjuncts (e.g., *Instruments* and *Locations*). Furthermore, these fixations to the object were anticipatory, occurring before the utterance of the noun referent (between 500 and 1000 ms after verb onset). Because these fixations did not occur more frequently in the typical than the non-typical condition for verb arguments, it seems that this advantage is specific to the argument structure encoded by the verb rather than world knowledge about typical verb-*Recipient* pairs.

However, in a second experiment, where there were two possible referents to the verb, the visual-attentional system preferentially relied on typicality over verb complement structure. In other words, there was a greater proportion of fixations to the object that was more typical even if it was an adjunct and not an argument. In this case, world knowledge took precedence over the information encoded by verb argument structure. Again, these fixations were anticipatory, with the advantage of typicality occurring within 300 ms following the verb's onset. Taken together with the results of the previous experiment, these findings indicate that visual information depicted in the array is integrated with verb-specific information in the selection of potential verb referents.

While Boland's (2005) results show a preference for world knowledge of typical verb participants over argument structure, the question of whether typical stored knowledge is preferred over knowledge activated by the depicted scene has also been examined. Knoeferle and Crocker (2006) presented participants with an ersatz scene consisting of clipart depictions of three characterized human figures (e.g., a wizard, a pilot and a detective). Two of the figures were referred to in the sentence, which took the German Object-Verb-Subject form, and could either take the role of a typical agent/patient (*The pilot [Patient] jinxes soon the wizard [Agent]*—or in the SVO construction, The wizard will soon jinx the pilot) or atypical agent/patient (*The pilot [Patient] jinxes soon the detective [Agent]*—or, The detective will soon jinx the wizard). Each scene depicted the two figures as either performing the given action (unambiguous) or not (ambiguous). Results showed that individuals preferred depicted events over stored thematic knowledge of typical verb participants, even when those depicted events were atypical. In other words, when the depicted scene was unambiguous, the target of the verb (the agent) was fixated more quickly than when the scene was ambiguous, even when that target agent was not stereotypical (e.g., a jinxing detective).

In a second series of studies, results showed that this preference was maintained even when the scene was not co-present during the unfolding of the spoken sentence (Knoeferle & Crocker, 2007). These results corroborate previous findings of the "blank screen paradigm" (Altmann, 2004), which showed that after an initial inspection of the scene that disappeared prior to the utterance, anticipatory eye movements were made to the location where the object had been on the screen. Knoeferle and Crocker (2007) take these results as implying that situated language comprehension makes use of a working

memory store in which entities/events within the scene inform the incremental processing of the unfolding utterance, and that the activation of these entities is resistant to decay for a certain duration during the processing of the sentence.

In sum, the main focus of visual world studies that rely on the information encoded by verbs, particularly typical role fillers (based on world knowledge), has been how the selectional restrictions of these verbs guide eye movements to their visual referents. Work by Altmann and Kamide (Altmann & Kamide, 1999; Kamide et al., 2003) suggests that the semantic constraints encoded by verbs can be used to predict possible role fillers even before their utterance. In addition, Boland (2005) has shown the syntactic properties of verbs, and their interaction with real world knowledge, also serve to trigger anticipatory eye movements toward pictures of objects. Knoeferle and Crocker (2007) have suggested that while thematic role knowledge does influence eye movement behaviour, depicted events are more salient even if they are inconsistent with that world knowledge. Finally, these depicted entities can draw fixations to their prior location even once they are removed. Taken together, the thematic roles encoded by verbs have consistently been shown to lead to fast, even anticipatory, fixations to their visual referents.

However, the "scenes" used in these studies consisted of *ersatz* scenes, which are missing some of the elements that constitute real-world scenes (Henderson & Ferreira, 2004). *Ersatz* scenes typically include the use of line drawings or clipart against a plain or low-feature background. This is in contrast to some studies that have used embodied contexts in which participants interact with objects in their environment (e.g., Spivey et al., 2001). Notably missing from the studies used in the visual world paradigm is the

middle ground between *ersatz* scenes and the real world—i.e., real motion in complex naturalistic scenes. In our previous work (Di Nardo, 2005) upon which the studies presented here are based, we used such real-world scenes.

Verbs and Dynamic Scenes in the Visual World Paradigm

To our knowledge, Di Nardo (2005; see also van de Velde, 2008) has been the first study investigating language comprehension in realistic, dynamic real world scenes (films of events). This manipulation increases the ecological validity of the study of situated language processing without introducing the task demands often present in studies of the embodied visual world. In particular, it allows for the construction of event meaning based both on the unfolding utterance and the dynamic scenes. These realistic scenes are feature-rich and complex, and may consume a greater proportion of the visual-attentional system than the linguistic stream, thus leading to longer delays between the onset of linguistic stimuli and the initiation of eye movements that reflect the processing of such stimuli. While a scene's gist can be extracted very quickly, usually within 100-300 ms (e.g., Potter, 1976; Intraub, 1999), those that are dynamic may include shifting "gists," or meanings, as various events unfold. Therefore, in order to continually construct a scene event's meaning, the primacy of the dynamic elements of a given scene may command a larger portion of the visual-attentional system. This should lead to increased attention to the most meaningful elements of the scene, particularly those "in motion," such as human figures or moving objects (e.g., De Graef, 1998). In particular, some results have shown that human faces automatically attract fixations (Morand, Grosbras, Caldara, & Harvey, 2010), and that their eyes direct attention to the targets of their gaze pictured in the scenes (Weith, Castelhano, & Henderson, 2003).

In order to further examine the role of verb information in the interaction between linguistic and visual representations, Di Nardo (2005) conducted two experiments that employed realistic scenes (both pictures and movies of events). In that study, we tested the hypothesis that the interaction occurs at a post-modular, conceptual level. Based on Dowty's (1991) notion that thematic role information is conceptual in nature, we hypothesized that the interaction occurred beyond the initial parsing conducted by the linguistic system, but within a working memory system termed conceptual short-term memory (CSTM; Potter, 1999; see also Potter, 1993), which integrates the outputs of both visual and linguistic input systems.

Eye movements were recorded as participants listened to sentences referring to an event about to take place in the immediate future, while viewing scenes related to those events. The sentences we employed contrasted verbs from two different classes: causative and perception verbs (as classified by Levin, 1993). The thematic role pairs encoded by each verb class are different: by hypothesis, Causative verbs denote an entity (the *Patient*) that undergoes a change of state caused by an *Agent* (the "doer" of the action denoted by the verb). On the other hand, perception verbs involve a *Theme* that does not undergo a change of state but is rather the *stimulus* perceived by an *Experiencer* or the *causer* of a given psychological state (e.g., *notice*). Given this difference in verb structure and meaning, one would expect causative verbs, which are selectionally more restrictive, to guide visual attention toward the referents of the objects they refer to more quickly than perception verbs. In addition, perception verbs are not only less restrictive, but also place a heavier emphasis on the human figure within the scene (the *Experiencer*) thus shifting visual attention away from the object (*Theme*).

In the first experiment of Di Nardo's (2005) work, the scenes were static, consisting of a still frame taken from the movies used in the second experiment. Two variables were manipulated: first, the verb used in the main clause of the sentence, which was either a causative (such as *The woman will crack the eggs that are on the counter*) or a perception verb (such as *The woman will inspect the eggs that are on the counter*). In addition, we manipulated the direction of motion taken by the agent present in the scene (actual motion in the second experiment, and "implied" motion in the first, evidenced by the agent's unambiguous orientation with respect to the target object in the scene). The agent either appeared to be (or was) moving toward the target object named in the main clause, away from it, or neither (the "neutral" condition). In the film clips, the onset of the verb was synchronized with the onset of motion (or the equivalent time point in the neutral films); this was termed the disambiguating point.

We hypothesized that if eye movement behaviour is locked to the perceptual and cognitive processes involved in situated language processing, then saccades would be launched toward the target object more quickly with causative than perception verbs, as causative verbs are more semantically restrictive and thus constrain the number of possible referents in the scene to one (the target object). In addition, we hypothesized that saccades would also be launched more quickly when the agent either appeared to be moving, or actually did move, toward the target object (as opposed to away from it or neither). Because the motion context in the toward condition visually confirmed the intended agent-object interaction, and demonstrates a tight semantic relationship, we expected eye movements toward the target object to be particularly fast in the causative-toward condition.

Results from the first experiment using static scenes demonstrated that only motion context had a significant main effect, such that saccade onset times (SOTs) were shorter when agents appeared to be moving toward the target object than when they appeared to be moving away from it or remained in a neutral position. The contrast between the two verb classes failed to produce a significant difference in SOTs when analysed by participants, although a marginally significant effect was found by items. In addition, overall SOTs were relatively high compared with the results of the Altmann and Kamide (1999) study, and did not show any anticipatory effects, with saccades being launched 486 ms *after* the offset of the noun. Interestingly, however, when we compared the effect of verb type in the toward condition, a significant difference was found (only in the item analysis). In other words, when the motion context was consistent with the agent-object interaction implied by the verb, saccades were launched more quickly toward the named object. These contrasts did not hold true in either of the other two motion contexts.

These findings failed to support the notion that verb-specific information, in the form of thematic or conceptual roles, can guide eye movements, which is contrary to the bulk of the findings reported in the literature (Altmann & Kamide, 1999; Kamide et al., 2003; Boland, 2005). Much stronger support was shown for the effect of the apparent motion of the agent on eye movements, a variable not previously examined in the literature. It appears that when realistic depictions of everyday scenes are employed, the visual context takes precedence over the linguistic context in constraining the domain of subsequent reference. Specifically, information conveyed by the scene, while static in

nature, was sufficient to establish reference to the target object, while any possible effects of verb class are not.

To determine whether these effects would hold for a dynamic visual context, Di Nardo's (2005) Experiment 2 employed dynamic scenes. Because situated language comprehension occurs mainly within a dynamic visual world, this experiment served to increase the ecological validity the first experiment. It was expected that causative verbs to lead to shorter SOTs than perception verbs; and for the condition in which the agent moves toward the garget object, shorter SOTs than the neutral and away conditions were expected. However, results with dynamic scenes yielded a similar pattern of results as in Experiment 1: there was a significant main effect of motion context, with the toward condition having a shorter SOT than the away condition, which in turn had a shorter SOT than the neutral condition. As in the previous experiment, verb type did have a significant effect, but only in the toward condition. Eye movements were on average launched 211 ms after the offset of the noun, which were faster than in the static scenes, but did not constitute evidence for anticipatory eye movements.

On the whole, these results failed to show that linguistic information contained by the verb could constrain the domain of reference when scenes are visually complex as they are in real world environments. We took the relatively late saccade onset times to be evidence for the modular (i.e., independent) processing of the linguistic and visual systems, and proposed that whenever interactions do occur they happen at a post-perceptual level, something akin to the conceptual short-term memory system proposed by Potter (1999). However, the time course of these eye movements also raises an important question about the study of verbs in dynamic, realistic scenes: is the control for

visual fixation truly *anticipatory*, or does the visual-attentional system use scene information to *confirm* unfolding event representations?

Building Event Meanings Through Confirmatory Eye Movements

The basic premise of the visual world paradigm is that eye movements are closely tied to the online cognitive processes of language comprehension. Indeed, saccadic eye movements are the means by which selective visual attention is manifested, and reflect visual search processes that indicate what people find most salient (i.e., most important and meaningful) about their visual environment (Henderson & Ferreira, 2004). A number of visual world studies have shown that the eye movements to objects of interest often occur shortly after their utterance (e.g., Allopenna, Magnuson, & Tanenhaus, 1998). In some cases, when the preceding utterance is predictive of a given entity (such as the visual representation of a stereotypical grammatical object of the uttered verb), eye movements may even be anticipatory, occurring before the noun is uttered, as in the Altmann and Kamide (1999) study.

The results of these studies are often construed as evidence for the anticipatory search for appropriate targets to the preceding utterance. However, we hypothesize that these eye movements serve instead as confirmatory of the current (albeit temporary) representation of event meaning. The temporal occurrence of these eye movements is of particular importance—those initiated before the utterance of the noun are taken to be confirming likely targets of *the thematic role to be filled*, while those occurring after the noun's utterance confirm the *presence of the entity in the visual scene*. The crucial distinction here is the source of the representation to be confirmed: one has been

activated by thematic knowledge stored within the verb, while the other has been activated by the presence of entities contained within the scene.

Whether visual search is initiated before or after the noun's utterance is likely influenced by the visual complexity of the scene. The more salient the visual context (whether because it contains motion, or because it contains highly informative elements), the more likely it is to contribute to the interpretation of event meanings. As Di Nardo (2005) suggested, the relatively "late" eye movements can be interpreted as confirming the event meaning conveyed by the unfolding utterance; in that study, SOTs were fastest in the condition where the motion context (toward) was consistent with the semantic constraints of the verb (causative).

In addition, the time course in which saccades are initiated is not always as rapid as the so-called "express saccade" (taken to be approximately within 200 ms of the stimulus presentation) which only applies when the attentional system is disengaged. When it is engaged, longer times are required to initiate a new saccade because the attentional system must disengage from the currently fixated location (Fischer & Breitmeyer, 1987). Therefore, in visual world studies that report fast and automatic attentional shifts, it is important to consider that the low visual complexity of the static *ersatz* scenes usually employed permits the attentional system to disengage quickly. In such scenes (e.g., Altmann & Kamide, 1999), the low complexity allows for the scene's few objects to be simultaneously held in working memory, producing rapid eye movements to likely candidates of thematic roles. Thus, when the more semantically restrictive verb was uttered, the appropriate object was sought out by the visual-attentional system to confirm the event meaning being constructed, and shifts in visual

attention were more likely to be influenced by linguistic constraints. In addition, because the scene is not likely to change as it is in our dynamic scenes, there is more reason to attend to properties of the linguistic utterance, without having to divide attention between two sources of potential information. In realistic and dynamic scenes, however, the primacy of the agent's presence and motion likely serve as the prime candidate for valuable information as to the events about to take place. With these scenes (such as those used by Di Nardo, 2005), the salience of the visual context may exceed that of the linguistic context. It is therefore also important to determine how the features of these scenes may continually engage the visual attention system and affect the pattern of eye movements.

In addition to the relative poverty of scenes used in the visual world paradigm, the types of sentences employed may have also prompted eye movements to occur before the utterance of the noun referent. In sentences that impose a task demand, eye movements are likely to take place before the noun's utterance because of the necessity to initiate a motor sequence; participants are likely to rely on visual cues in as rapid a manner as possible to respond to the task demand. These task demands (i.e., to manipulate an object; e.g., Spivey et al., 2001) create a cognitive set in which likely targets of the task in the visual scene are actively sought out, thus influencing both the time course and scan path of eye movements. Indeed, in the two experiments conducted by Altmann and Kamide (1999), longer SOTs were found when there was no task demand. This suggests that the control of visual fixation is partially mediated by the cognitive set created by the experimental task encoded by the utterance; in more passive linguistic conditions, visual

search is likely to be a function of the more salient scene features rather than any cognitive set created by the spoken language.

In summary, the bulk of the literature examining the interaction of the visual and linguistic systems have argued for incremental (and non-modular) processing on the basis of the speed at which visual representations are fixated in response to the syntactic and semantic constraints imposed by verbs. These so-called anticipatory eye movements, while fast, were by no means obligatory (i.e., did not automatically occur in every, or even most, trials). In addition, several features of both the visual and linguistic contexts likely contributed to the speed at which these eye movements occurred. The failure of our previous work (Di Nardo, 2005) to replicate such findings (both the anticipatory eye movements, and the verb effects) in static and dynamic scenes likely reflects how the linguistic processor interprets language in realistic contexts. Given the primacy of the visual context in controlling visual attention, it may be that visual attention was being used to confirm the participation of the unfolding event's role fillers. In other words, the event meaning being constructed was more heavily reliant on the salient visual features of the scene than the linguistic utterance, and eye movements were rendered insensitive to verb effects.

Verbs and Dynamic Scenes: The Visual Grounding of Event Meanings

Because verbs lexicalize events, which are usually visually represented as unfolding, dynamic "happenings" (in Levin and Rappaport Hovav's sense of the word, 1998), the true grounding of the event meanings conveyed by verbs ought to be grounded in dynamic scenes. However, as discussed above, no studies (other than Di Nardo, 2005 and van de Velde, 2008) of situated language comprehension have taken place within

dynamic scenes. It is likely that this lack of dynamic context, combined with the types of sentences and task sets used, has influenced the pattern of results supporting verb-guided anticipatory eye movements. In addition, the question of whether eye movements are anticipatory or confirmatory still remains open.

The purpose of the present study is to investigate the discrepant results found in the visual world literature focusing on verbs. While studies with static—or *ersatz*—scenes find anticipatory eye movements to verb-thematically related objects in the scene, Di Nardo (2005) found no such effects using dynamic scenes. We aim to explore how manipulating the conditions in which situated language processing takes place, by emphasizing either aspects of the visual or linguistic context, guides visual attention. In addition, a larger goal of this research is to study how events grounded in realistic, dynamic visual contexts aids in the incremental construction of event interpretations unfolding across time and space.

Overview of the Present Research

There are three key issues that the present research aims to address in order to examine why Di Nardo (2005), which was the first study to employ realistic static and dynamic scenes, failed to produce consistent verb effects. Given this relative insensitivity to verb class found, methodological changes to the visual world paradigm were introduced. In addition, we also addressed the question of whether eye movements serve to anticipate or confirm likely candidate role fillers, particularly in the context of dynamic scenes.

The first key issue attempted to determine whether the lack of verb effects in Di Nardo (2005) were a function of covert attention shifts without accompanying eye

movements. Our previous research has shown that the relationship between visual attention and the linguistic system is not automatic nor necessary (Di Nardo, 2005), and shifts of attention can be accomplished without a corresponding shift in gaze (Fischer & Breitmeyer, 1987). That is, it is possible that covert attention to objects without foveation might have occurred allowing for the appropriate thematically related object to be integrated with event interpretation, thus suppressing anticipatory eye movements commonly found in static scenes. Because the movies in Di Nardo (2005) were presented on a computer screen subtending 50.4 degrees of visual arc, we hypothesized that attentional shifts could have occurred without corresponding eye movements. In order to "force" saccades towards the various entities in the scene, in Experiment 1 we increased the visual angle of the scene by projecting it onto a large canvas screen subtending 70.4 degrees. By introducing this manipulation, we hypothesized that the attentional system would be more easily disengaged from its current fixation location and be influenced by the linguistic properties of the utterance. We expected that not only would SOTs be shorter, but that our measures would be more sensitive to verb contrasts.

The second key issue addressed whether the lack of verb effects from previous studies are a function of lack of attention to linguistic properties of utterances. In the second experiment, we explored whether the absence of the linguistic context would alter the pattern of eye movements. We thus contrasted two conditions; one in which the films were presented without accompanying sentences, and one with the sentences (what could be considered the inverse of the "blank screen paradigm;" Altmann & Kamide, 2004). We expected that the presence of the linguistic context would lead to faster, and more, fixations to the target object after the disambiguating point.

The third key issue examined whether a semantically stronger linguistic context could drive verb-mediated saccades to target objects. The third experiment strengthened the effect of the linguistic context by altering the nature of the initial patch clause used in each of the sentences. We contrasted two clauses, one of which was semantically restrictive, or predictive of the event referred to in the main clause (e.g., *In order to make the omelette...*), and another which was non-restrictive (e.g., *After pouring the flour into the bowl...*), prior to the main clause containing either a causative or an *experiencer* verb (*...the woman will crack/inspect the egg*). We hypothesized that the more restrictive clause would lead to shorter SOTs as it contributes more strongly to the building of the event meaning encoded by the main clause.

Experiment 1

The main purpose of this experiment was to replicate the Experiment 2 of Di Nardo (2005) with the same sentences and dynamic scenes, but under different experimental conditions. More specifically, instead of presenting the visual stimuli on a computer monitor, we projected the film clips on a large screen, thus increasing the scene's visual angle from 50.4 to 70.4 degrees of visual arc. This was done in order to preclude the possibility of visual-attentional shifts without corresponding eye movements (Fischer & Breitmeyer, 1987). The study was conducted with the aim of investigating previous findings (e.g., Altmann & Kamide, 1999) supporting the idea that eye movements are driven by linguistic constraints (in this case, verb-specific information). We expected the methodological change we introduced to lead to a significant main effect of verb type. In addition, this experiment aimed to replicate the finding that the direction of motion of agents in these scenes would have an effect on the time course of

post-verbal eye movements. As in the previous study (Experiment 2, Di Nardo, 2005), we expected the toward motion condition—i.e., when the agent moves toward the referent of the grammatical object of the main verb—to have shorter SOTs than the away and neutral conditions.

Because the visual world literature often points to frequent anticipatory eye movements (e.g., Allopenna et al., 1998; Altmann & Kamide, 1999), we also examined very early post-verb eye movement behaviour to test for possible anticipatory effects. Contrary to the findings of previous studies, we did not expect to find any such effects, as the complex scene features (including motion) should have primary control of visual attention, causing eye movements to be confirmatory rather than anticipatory (as was found in Di Nardo, 2005). However, we did expect verb type and motion type to begin showing main effects on the number of saccades made to the target object prior to the offset of the noun.

Finally, the relationship between target object and event saliency, and eye movement behaviour, was examined. Based on norms obtained in our previous study (Di Nardo, 2005), we determined how salient the target object was in each scene. In some cases, the scenes have rather complex backgrounds and several competing objects of interest. In addition, we also calculated how salient, or "predictive," each scene was in terms of the event described in the sentence. We expected that these saliency ratings would correlate with longer fixation times to the target object, shorter SOTs, and lead to a higher likelihood of fixating the target object at verb-onset.

Method

Participants. Forty-four participants took part in this study, all drawn from the Concordia University student body. None of these participants had taken part in the earlier experiments. There were 32 females and 12 males, ranging in age from 18 to 40. Data from 32 of these participants were retained in the analyses, the rest having to be discarded due to computer calibration or recording errors. Inclusion criteria for participation included being a native speaker of English (defined as having learned the language by the age of five), and having normal or corrected-to-normal vision with contact lenses (participants wearing glasses were excluded due to interference with the eye-tracker). All participants received course credit for their participation.

Materials and apparatus.

Stimuli. The stimuli used in this experiment were identical to those used in Experiment 2 of Di Nardo (2005), in order to keep the experimental conditions constant while varying the size of the visual angle of the scenes. There were a total of 102 sentence/movie combinations. The same structure was employed for each of the 17 sentence pairs (see Appendix A for complete list of the stimulus sentences and their corresponding visual scenes): there was an initial patch clause (always of an adverbial type), a main clause, and a second patch clause. For example, in the sentence, *Before making the dessert, the cook will crack the egg that is in the bowl,* the main clause is *the cook will crack the egg,* and the two patch clauses are *Before making the dessert* and *that is in the bowl.* The main clause always took the form of NP1 (Noun Phrase 1)-*will*-Verb-NP2-RC (Relative Clause). The NP1 always made reference to the agent in generic form (e.g., *the cook, the boy,* etc.); the NP2 (the direct object of the verb) made reference to the

target object in the scene (e.g., *the egg*); and the RC always made reference to the target object (e.g., *that is in the bowl*).

The 17 verb pairs in each main clause were selected from two verb classes, based on Levin's (1993) classification; causative or perception/psychological verbs (which will be referred to as "perception" verbs from this point forward). These verbs were matched based on frequency (Kucera & Francis, 1967) but also based on the plausibility of the events to be filmed (so as to allow for pairs such as *crack/examine the eggs*), as judged by the experimenters. Sentences were digitally recorded using SoundEdit 16 for Mac at 44.100 Hz by a female research assistant speaking at a normal pace.

The film clips were of naturalistic indoor and outdoor scenes, and were altered only to create a resemblance to common household scenarios, and to ensure the proper placement of the target objects. Agents and target objects were always on the same plane (mid-ground) at opposite sides of the screen (e.g., if the agent was on the left, the target object was near the right edge of the scene). There was no camera movement or zoom, and the only source of motion within the movies was that of the agent performing a given action. There were three versions of each of the 17 unique scenes (e.g., someone cooking in a kitchen): after an initial similar segment of about 7 s, agents moved (or reached) either towards a particular target object, away from it, or remained neutral (i.e., continued doing what they were doing in the initial segment; see Figure 1). There were thus a total of 51 unique movies (17 scenes x 3 endings). Film resolution was set at 720 x 480 pixels in NTSC format (29.97 frames per second), and each film lasted approximately 10 seconds.

Figure 1. Position of the agent in the three motion conditions of the same event at the disambiguating point. The frame is from a neutral motion condition film (i.e., when the agent—the cook—does not move towards or away from the target object—the eggs). The outlines represent the positions of the same agent in the corresponding frame of the movie in the towards motion condition (red), and in the away condition (blue). The scene and the outlines exemplify the different positions of the agent at the acoustic onset point of the two corresponding verbs (*crack* or *examine*), demonstrating the similarity between event onsets in the three versions of the movie.

Each of the three movies was then combined with each of the two sentences (causative and perception), which created 102 film/sentence combinations. The onset of the verb's utterance was synchronized with the onset of the agent's motion, which was defined by determining the frame in which the agent started moving towards or away from the target object (e.g., the start of rotation of the torso or limbs towards or away from target object), or the equivalent time point in the neutral movies. The acoustic onsets of the verb was determined by identifying the lowest frequency between words in the digital sound files, or by splitting transitional phonemes when the lowest frequency could not be clearly determined. The digital movies were produced and edited by a film student using FinalCut Pro (Apple, Inc.). The 102 stimuli were distributed in six lists of materials, each one containing 17 trials (film/sentence combinations), with two or three of each verb-type/motion-type combinations.

Apparatus. The film clips were projected by a Sharp XR-20X DLP projector onto a 114 cm by 160 cm blank canvas screen placed 236 cm in front of the participant. This created a visual angle of 70.4° of arc. This is in contrast to the previous study (Di Nardo, 2005, Experiment 2), where participants viewed the scene on a 21" computer monitor at a distance of 41 cm, subtending a visual angle of 50.4°. Thus, the present manipulation represents to a total increase of 71.6% of visual angle.

Participants wore head-mounted earphones through which the sentences were presented binaurally. The experiment was run with iQTrack software (custom-programmed software designed to run film clips and allow for the films to be superimposed on the eye movement paths in the post-experimental coding and processing software) on a PC with a Pentium 4 CPU running at 2.8 gHz with 1.5 GB of RAM.

Participants' eye movements were recorded using the ViewPoint PC60 EyeTracker (Arrington Research) at a sampling rate of 30 Hz from the right eye only (viewing was binocular). This device is mounted to a chinrest, designed to minimize head movements.

Procedure. Prior to the experiment, participants gave their informed consent and were given written instructions regarding the study (see Appendix B). Participants were assigned to one of the six experimental lists in consecutive order. The first phase of the experiment consisted of the manual calibration of the eye-tracker, which lasted approximately 10 to 15 minutes. Participants were then shown a short version of the instructions on screen, reminding them of how to proceed during the experiment (see Appendix C).

Participants were asked to watch the movies and to pay attention to the sentences, as a short recall task would be given afterward to test their memory of the films. Each trial began with a red fixation cross on a black background that appeared on the screen for two seconds. The fixation cross was included to orient participants' eyes at the beginning of each trial and to ensure a uniform starting point for all participants. This was followed by the first frame of the film clip with the fixation cross continuing to appear in the centre of the screen for one more second. At that point, the movie began and the acoustic onset of sentence began within a few seconds from there. Each trial lasted approximately 12 s. The 17 trials were presented in random order, with a few seconds in between each, during which time the screen appeared black.

The calibration and experimental phases were conducted in a dark room to minimize glare for the eye-tracking camera and to help participants focus their attention on the screen. At the end of the experiment, participants were given a short quiz, which

consisted of a 12-item cued recognition task containing six pictures and six written sentences, half of which were foils and the other half of which were taken from the experimental stimuli (the pictures were frames taken from the films). This task was to ensure that participants were paying attention during the experiment. The entire experiment lasted approximately 30 minutes.

Analyses. Four sets of analyses were conducted, the details of which will be described below in the Results and Discussion section. In brief, the first set examined the nature of the very early post-verbal eye movement patterns (between verb-onset and the offset of the noun), and the effects of verb type (causative *vs.* perception) and the agent's motion type (towards, away or neutral with respect to the target object) on these eye movements. Second, the effects of target object and target event saliency on overall eye movement behaviour were examined. The third set constituted the main analyses, wherein the effects of verb type and motion type on post-verbal eye movements were investigated. Finally, we compared the results of this experiment to the data obtained in the Di Nardo (2005) work upon which these studies are based. An alpha level of .05 was used for all statistical tests, unless otherwise indicated.

Results and Discussion

To ensure participants paid attention to the movie clips during the experimental trials, a short cued recall test was given at the end. Participants with scores less than 50% would not have been included in the main analyses. Quiz scores ranged from 83.3% to 100%; therefore, all participants were retained in the analyses.

A manipulation check also examined the effect of verb type (causative *vs.* perception) and motion type (away, neutral and toward) on the proportion of trials (by

items) where participants looked at the target object before, or at, the onset of the verb. We did not expect a difference between the six conditions as up to the disambiguating point all six versions of each movie (motion context and initial sentence segments) were similar. As expected, neither verb type nor motion type had an effect on the proportion of trials in which the participant initiated a fixation to the target object prior to the disambiguating point ($p > .05$).

Missing data. The proportions of missing data from three different sources were computed. The first source of missing data was due to corrupt data, such as a system crash, poor calibration or, more frequently encountered, drift caused by head movements. Out of the 544 trials presented to participants, 86 (15.8%) were lost due to corrupted data. These trials were distributed evenly across the various experimental conditions (no significant main effects or interactions).

The second source of missing data was trials in which participants never fixated the target object after verb-onset (note that these saccades are by no means obligatory, thus these trials are not true "missing data" in the strictest sense of the word). One hundred and eight (19.8%) such trials were recorded. This proportion is somewhat higher than that found in Altmann and Kamide's (1999) study, where participants never fixated the target object in 7% of trials.

In order to examine whether verb type (causative *vs.* perception) and motion type (away, neutral and toward) had an effect on the proportion of trials where participants did not launch a saccade to the target object after verb-onset, a repeated-measures 2 X 3 ANOVA ($N = 32$) was conducted (computed by participants). We expected that both verb type and motion type would have a significant main effect, such that there would be

fewer such missing trials in both the causative and toward conditions. Results showed that while verb type did not have a significant main effect, $F(1, 31) < 1$, $p = .54$, motion type did, $F(2, 62) = 11.34$, $p < .0001$. This indicates that while verb class did not affect whether participants did not fixate the target object after verb-onset, contrary to our expectation, the apparent motion of the agents in the scenes did influence the frequency with which the target object was not fixated after the verb-onset. There was no significant interaction effect between the two independent variables, $F(2, 62) < 1$, $p = .74$.

To further elucidate this finding, a post-hoc analysis (Scheffé's test) was conducted, which indicated that there was a significantly higher proportion of trials in which the target object was never fixated after the verb-onset in the away condition compared to the neutral condition ($MD = .096$, $p = .03$) and the toward condition ($MD = .169$, $p < .0001$). The comparison between the neutral and toward conditions was not significant ($MD = .073$, $p > .05$). This partially confirmed our hypothesis that the toward condition would have fewer such trials, and indicates that participants were less likely to launch a saccade toward the target object in the away condition than in the other two conditions.

One explanation for this finding may be that because the agent's motion did not conform to the expected path as specified by the utterance (i.e., toward the target object) in the away condition, participants were less likely to fixate the target object, and tended to stay fixated on the agent for the remainder of the movie. In fact, we found that most of the saccades and brief fixations during the initial linguistic processing of the event (between the onset of the verb and the offset of the noun) were to the agent. This observation is consistent with previous studies on static scene processing without

linguistic stimuli, which have found that visual attention is primarily directed towards human figures when they are present in a scene (Henderson & Ferreira, 2004). This is also consistent with findings by Kamide, Altmann and Haywood (2003) and Boland (2005) who found that animate agents attracted the most fixations in a scene. In addition, the results are also consistent with classic eye-tracking studies which showed that in paintings, fixations are predominantly made to human figures (Buswell, 1935; Yarbus, 1967).

In order to determine why participants might never have looked at the target object after verb-onset, a Chi-Square test was conducted to examine the effect of whether having looked at the target object before the verb was uttered affected post-verb eye movement behaviour. Results indicated no significant effect, $\chi = 2.02$, $p = .16$. Thus, it appears that previous encoding of the target object does not preclude later fixations, particularly those following the utterance of the verb or the noun itself.

The third source of missing data derived from trials in which participants happened to be fixating the target object at verb-onset. This occurred in 56 (10.3%) of the trials. These trials had to be excluded from any analyses examining the effect of verb type on subsequent eye movement behaviour, because of the inability to discern whether participants continued to fixate the object because they heard the more semantically-restrictive verb that could have involved that object or not. A 2 X 3 repeated-measures ANOVA ($N = 32$) was conducted to examine the effect of verb type and motion type on the proportion of these trials (by participants). We did not expect any significant main effects or interactions, as movies (and sentences) were essentially identical up the disambiguating point. As predicted, the results indicated that there was no significant

interaction effect, $F(2, 62) < 1$, $p = .09$, nor was there any main effects of verb type, $F(1, 31) < 1$, $p = .95$, or motion type, $F(2, 62) < 1$, p .40. These results support the idea that the verb class of the sentence and the agent's direction of motion uttered did not affect whether participants were fixating the target object at the time the verb was spoken. Hence, it can be said that these trials of missing data were evenly distributed across all conditions.

In summary, the pattern of missing data attributable to two of the three sources (corrupt data trials and trials where the participant was fixating the target object at verb-onset) was randomly distributed across the six conditions. However, trials where participants never looked at the target object after verb-onset were not evenly distributed across the two verb conditions, such that fewer fixations were made to the target object in the away condition. Because such a large proportion of data was missing overall, the analyses reported below (both by participants [*F1*] and by items [*F2*]) had any missing cell means replaced with condition means.

Analysis of early post-verb cumulative saccades to the target object. This analysis is based on a combination of data analysis techniques suggested by Altmann and Kamide (2004). The purpose was to determine if verb-guided eye movements would be closely time-locked to the utterance of the verb, especially before the offset of the noun object, as has been shown in other studies (e.g., Altmann & Kamide, 1999). First, the cumulative number of saccades that were initiated towards the target object during each 50-ms bin following the onset of the verb was computed. In other words, for each trial, the number of saccades made in the first 50 ms after the verb-onset was counted, and so on for every 50-ms interval following that point.

Next, for each critical point in the sentence (verb-offset, noun-onset and noun-offset), the mean cumulative number of saccades to the target object was calculated for each condition. In other words, we counted how many saccades toward the target objects had been launched at each of these points in the main clause. Because the critical sentence points differed for each sentence, based on the length of the individual words for each as well as the speaker's rate of speech, the corresponding 50-ms bins were different for every sentence. Thus, the cumulative number of saccades for each participant was taken from these different bins. Despite occurring at different points in time, they corresponded to the same linguistic markers, namely the end points of the verb and noun, as well as the onset of the noun.

Finally, the effects of sentence point, verb type and motion type on the cumulative proportion of saccades to target object after verb-onset were examined using a 3 (sentence point: verb-offset, noun-onset, noun-offset) X 2 (verb type: causative *vs.* perception) X 3 (motion type: away, neutral, towards) repeated-measures ANOVA (both by participants, *F1*, and by items, *F2*). Trials in which participants were already fixating the target object at verb-onset were not included in the analyses.

We expected that there would be a significant interaction between sentence point and verb type, such that the difference between the two verb types would increase as the sentence unfolded, as well as a significant interaction between sentence point and motion type, such that the difference between the toward and away/neutral conditions would increase as the sentence unfolded. We hypothesized that the interpretation of the verb and the noun phrase should proceed incrementally, thus providing more restrictive information as time proceeds, especially in the more semantically restrictive causative

condition. This would indicate that verb-specific information can constrain visual attention at the very early stages of processing. We also expected there to be a significant effect of motion type.

The results (see Figures 2 and 3, as well as Tables 1 and 2) partially confirmed our hypotheses, such that the main effects of sentence point, $F1(2, 62) = 46.85$, $p < .0001$, $F2(2, 32) = 68.76$, $p < .0001$, and motion type, $F1(2, 62) = 4.74$, $p = .01$, $F2(2, 32) = 3.02$, $p = .06$, were significant. The interaction between sentence point and motion type was also significant, $F1(4, 124) = 6.13$, $p = .0002$, $F2(4, 64) = 4.96$, $p = .002$, while the interaction between sentence point and verb type was not, $F1(2, 62) < 1$, $p = .96$, $F2(2, 32) < 1$, $p = .78$. The three-way interaction of sentence point, verb type and motion type was not significant, $F1(4, 124) = 1.34$, $p = .26$, $F2(4, 64) < 1$, $p = .61$, nor was the two-way interaction between verb type and motion type, $F1(2, 62) = 2.45$, $p = .09$, $F2(2, 32) = 1$, $p = .26$, or the main effect of verb type, $F1(1, 31) < 1$, $p = .56$, $F2(1, 16) < 1$, $p = .43$.

To further explore the two-way interaction of sentence point and motion type, three tests of the simple effects of motion type across all levels of sentence point were conducted. The first two of these tests indicated that the effect of motion type was not significant at verb-offset, $F1(2, 62) < 1$, $p = .45$, $F2(2, 32) < 1$, $p = .55$, nor at noun-onset, $F1(2, 62) = 1.73$, $p = .19$, $F2(2, 32) = 1.24$, $p = .30$. However, at noun-offset, motion type did have a significant main effect, $F1(2, 62) = 7.05$, $p = .002$, $F2(2, 32) = 6.61$, $p = .004$. A modified Bonferroni/Dunn post-hoc analysis, with adjusted alpha levels of .03 for each of the three pairwise comparisons, showed that the toward condition had a significantly higher mean than the away condition, $MD1 = .231$, $p = .001$, $MD2 = .187$, $p = .002$, and neutral condition, $MD1 = .214$, $p = .003$, $MD2 = .155$, $p = .008$.

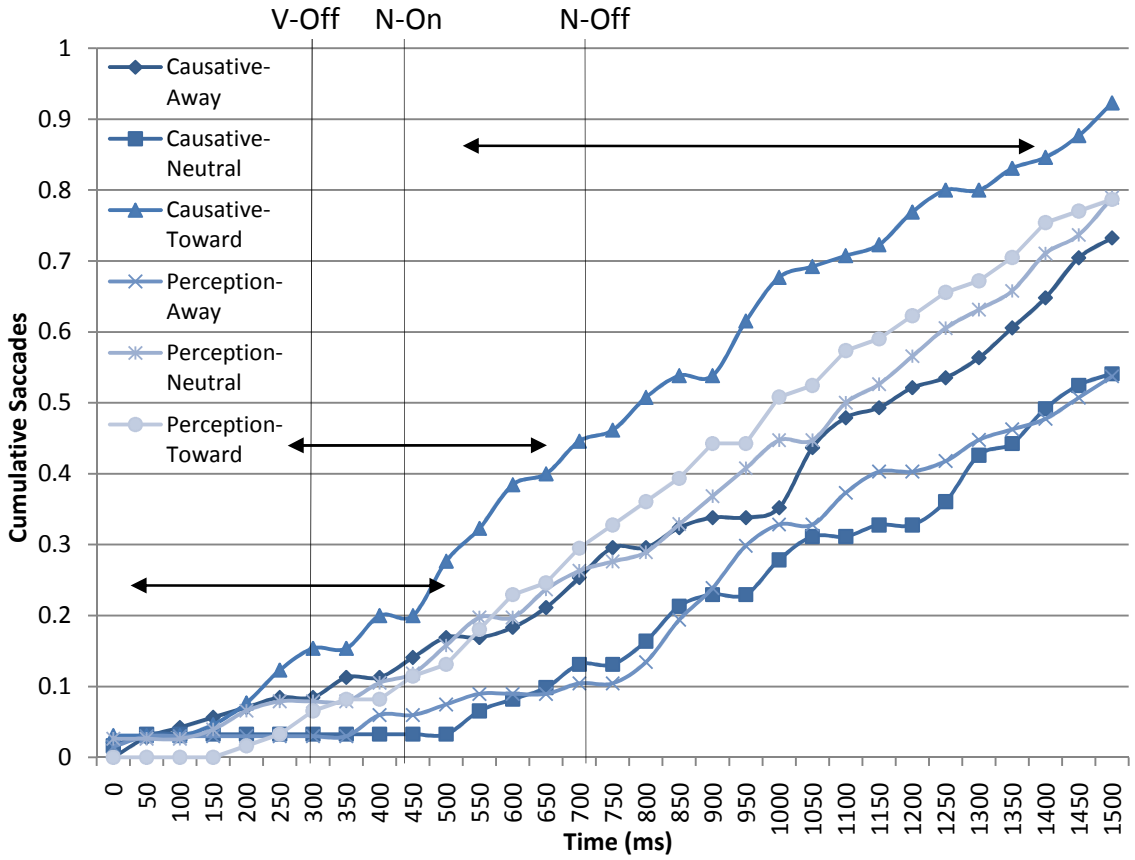


Figure 2. A plot of the mean cumulative average number of fixations (by participant) to the target object after verb-onset. Each line refers to a single condition, and each point to one 50-ms bin. The origin of the X-axis refers to the verb-onset, and the three vertical lines mark the temporal boundaries of the verb and noun (average onset and offset). The double-headed horizontal arrows on each boundary indicate the range of onsets and offsets of the different grammatical categories at the points in time relative to the verb-onset. These arrows take into account the variable lengths of each of the sentence segments (i.e. verbs and noun phrases). The point at which each of the coloured lines (referring to cumulative fixations for each condition) intersects with the three critical sentence points (verb-offset, noun-onset and noun-offset) were computed and compared in an ANOVA.

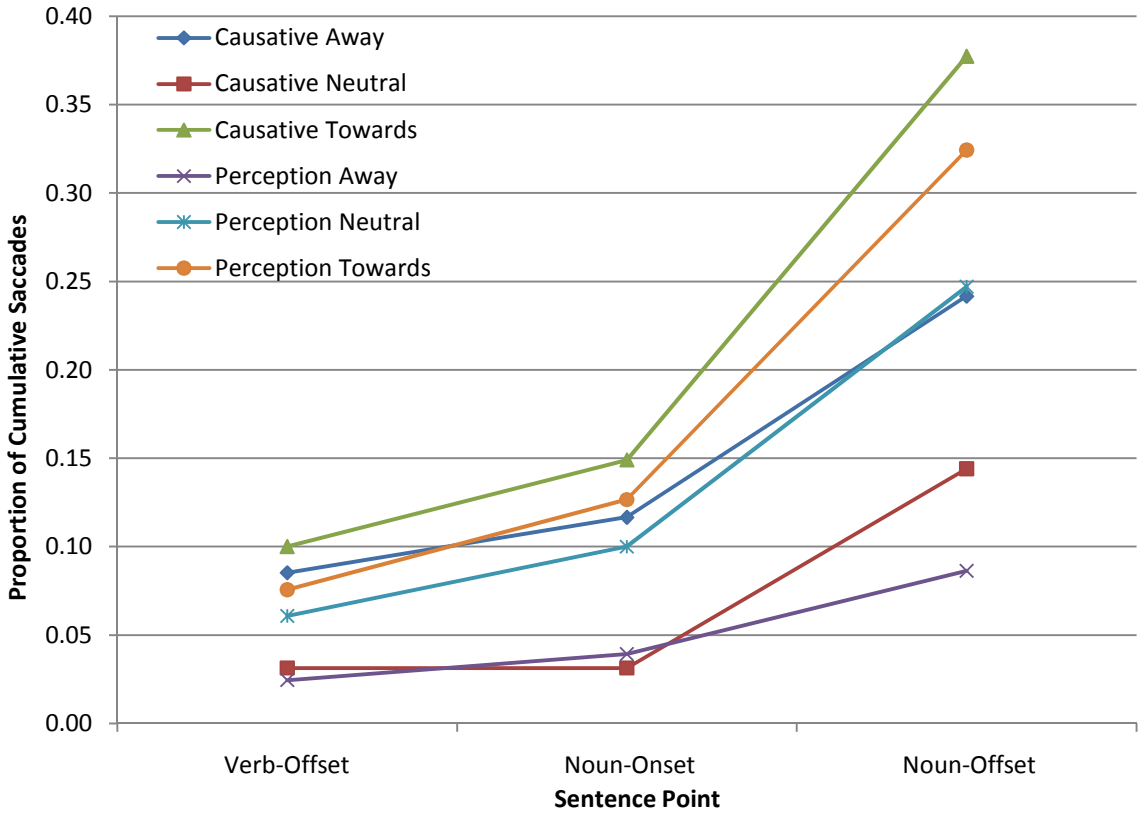


Figure 3. Mean number of cumulative saccades towards target object at each of the three critical sentence points, verb-offset, noun-onset and noun-offset. By the offset of the noun, the two toward conditions have the highest mean number of cumulative saccades.

Four planned comparisons, which consisted of a series of *t*-tests, were also conducted (by participants, *t1*, and by items, *t2*). The first two tested the hypotheses that the toward group would exhibit significantly higher means than the away and neutral groups across all sentence points. We expected that during the utterance of the event, the toward motion condition would lead to a higher number of cumulative saccades than the other two conditions. This hypothesis was supported: a one-tailed paired *t*-test indicated that the toward condition (*M1* = .199, *SD1* = .173; *M2* = .190, *SD2* = .121) had a significantly higher mean than away condition (*M1* = .099, *SD1* = .163; *M2* = .097, *SD2* = .128), *t1*(31) = 2.46, *p* = .01, *t2*(16) = 2.56, *p* = .01, and the neutral condition (*M1* = .103, *SD1* = .154, *M2* = .109, *SD2* = .095), *t1*(31) = -2.26, *p* = .02, *t2*(16) = -2.10, *p* = .03. These results indicate that, in fact, very early dynamic scene motion is incrementally integrated with the linguistic utterance of the unfolding event to prompt eye movements toward the object involved in that event.

Next, to examine the difference between causative and perception verbs without the moderating effect of apparent motion, we compared the causative-neutral and perception-neutral groups across all sentence points, with the expectation that causative-neutral would have a significantly higher mean than the perception-neutral groups, thus showing more pure verb semantic restriction effects. This hypothesis was not supported. Results indicated that the causative-neutral condition (*M1* = .061, *SD1* = .141, *M2* = .069, *SD2* = .101) in fact had a marginally significantly lower mean than the perception-neutral condition (*M1* = .136, *SD2* = .234, *M2* = .136, *SD2* = .139), *t1*(31) = -1.57, *p* = .06, *t2*(16) = -1.56, *p* = .07, contrary to our hypothesis. This indicates that in the absence of agent

motion, eye movements occurring immediately after the verb's utterance are not sensitive to the semantic restrictions encoded by the verb.

Finally, we compared the causative-towards and perception-towards groups, hypothesizing that the mean would be higher in the former than the latter. This is because we would expect early agent movement toward the target object, which is consistent with the event described by the main clause of the sentence, should enhance the sensitivity of the visual-attentional system to the semantic restrictions imposed by the verb. The hypothesis that the causative-towards condition (*M1* = .258, *SD1* = .276, *M2* = .209, *SD2* = .156) would have a higher mean than the perception-towards condition (*M1* = .146, *SD1* = .203, *M2* = .176, *SD2* = .246) across all sentence points was not supported, $t1(29) < 1$, $p = .17$, $t2(16) < 1$, $p = .68$ (the comparisons at each of the three sentence points were not significant either). Thus, it appears that at this early stage of post-verb processing, the presence of motion toward the target object does not lead causative verbs to engender to more saccades than perception verbs.

On the whole, the results of this set of planed comparisons point to the primacy of dynamic scene information in programming saccades soon after the disambiguating point. By the offset of the noun, verb type has no effect in restricting the domain of visual attention, contrary to our hypotheses, as well as the findings of previous studies (i.e., Altmann & Kamide, 1999). Taken together, these findings indicate that in the presence of dynamic scenes, event meanings tend to be rapidly influenced by the motion of the human figure. In this case, any semantic information encoded by the verb does not contribute to the building of event meaning.

While the motion context did have an effect on early post-verb saccades, this cannot be taken as evidence of anticipatory eye movements; even in the toward condition, there was only approximately a 20% probability that the target object would be fixated by the offset of the noun's utterance. The series of analyses reported below serve to address this question more directly.

Anticipatory eye movements. In order to determine whether any anticipatory eye movements may have been made, we calculated the proportion of trials in which a saccade was launched towards the target object before the onset of the noun, as well as by its offset. In addition, saccade onset times (SOTs**;** the time taken to initiate a saccade toward the target object after the onset of the verb's utterance) were compared to two other time points in the sentences: the noun-onset and the noun-offset. Reaction times were computed by subtracting the latency between verb-onset and noun-onset, as well as between verb-onset and noun-offset, then averaged by condition. This was done to examine how long it took eye movements to be initiated after hearing the verb but prior to hearing the onset and offset of the noun. This is in contrast to the previous set of analyses, which measured the *number* of saccades made during this time span, not the length of time the programming of these saccades required.

These differences in time between the onset and offset of the noun and SOT were computed and subjected to a 2 (verb type) X 3 (motion type) repeated-measures ANOVA, in order to determine whether these first saccades were affected by verb type and motion type. We predicted that there would be a main effect of both variables, such that the causative condition would yield a lower mean difference than the perception condition, and the toward condition would yield a lower mean difference than the away

and neutral conditions. But contrary to our prediction, no evidence of anticipatory eye movements was found. On average, eye movements were initiated 794 ms after the noun-onset in the causative condition, and 807 ms after the noun-onset in the perception condition. In addition, saccades were launched 534 ms after the offset of the noun in the causative condition, and 523 ms after the offset of the noun in the perception condition. In terms of the proportion of anticipatory eye movements, saccades were launched toward the target object before the onset of the noun in only 11.1% of the causative trials and 12.7% of the perception trials (the difference between these groups was not significant, $p > .05$). However, saccades were launched toward the target object before the offset of the noun in 25.4% of the causative trials and 27.6% of the perception trials (again, the difference between these groups was not significant, $p > .05$). Thus, it appears that although some eye movements were launched toward the noun referent by the end of its utterance, these saccades were by no means obligatory nor closely time-locked to the utterance. Again, this does not support the notion that the visual-attentional system is primarily used to anticipate objects involved in events described by the unfolding sentence, but is used rather in the service of confirming their presence once the event has been described, and agent movement unambiguously confirms this interpretation.

Next, the difference in time between the onset of the noun and SOT was computed and subjected to a 2 (verb type) X 3 (motion type) repeated-measures ANOVA, in order to determine whether these first saccades were affected by verb type and motion type. Results indicated that only the main effect of motion type was significant, $F1(2, 62) = 8.72$, $p = .0005$, $F2(2, 16) = 6.37$, $p = .005$. A modified Bonferroni/Dunn test, with corrected alpha levels of .03 for each of the three pairwaise

comparisons, was conducted to explore the significant main effect of motion type. This revealed that the toward condition (*M1* = 642.7, *SD1* = 670.1, *M2* = 580.7, *SD2* = 358.2) had a significantly lower mean than the neutral condition (*M1* = 1075.6, *SD1* = 719.8, *M2* = 946.2, *SD2* = 481.2), $p = .0005$, $p = .001$. The away group (*M1* = 895.4, *SD1* = 611.9, *M2* = 760.0, *SD2* = 432.6) did not significantly differ from either the neutral or the toward groups, $p > .03$.

This analysis was also conducted for the difference in time between the offset of the noun and SOT. Results indicated, again, that only motion type had a significant main effect, $F1(2, 62) = 9.91$, $p = .0002$, $F2(2, 32) = 6.35$, $p = .005$. A modified Bonferroni/Dunn test indicated that the toward condition (*M1* = 350.157, *SD1* = 685.6, *M2* = 302.4, *SD2* = 391.3) had a lower mean than the neutral condition (*M1* = 817.2, *SD1* = 724.3, *M2* = 668.4, *SD2* = 509.5), $p < .0001$, $p = .001$, as well as the away condition (*M1* = 627.8, *SD1* = 623.5, *M2* = 478.6, *SD2* = 476.7), but only in the participant analysis, $p = .01$, $p = .10$. The away condition did not significantly differ from the neutral condition, $p > .03$.

These results indicate that at both the onset and offset of the noun, semantically restrictive verb information does not preferentially lead to faster saccades of that noun's visual referent. Instead, as the results of the previous set of analyses showed, it is the information conveyed by the agent's direction of motion that affects the time course of eye movements toward the target object.

Given the primacy of event agent motion in affecting eye movement behaviour, we tested whether visual attention was locked to the agent's path of motion throughout the event. To that end, the difference between SOT and the time at which the agent

reached the object was computed. This would allow us to determine whether participants initiated a saccade towards the target object *before* the agent in the scene reached it (in the towards condition only, as the agent never interacted with the object in the away and neutral conditions). It was found that participants launched a saccade towards the target object 952 ms *before* the agent made contact with the object. In addition, a one-way repeated measures ANOVA revealed a significant main effect of verb type in the item analysis, $F2(1, 16) = 4.89$, $p = .04$, such that the target object was fixated more quickly in the causative condition than in the perception condition, but not in the participant analysis, $F1(1, 31) < 1$, $p = .93$. Thus, participants were not simply following the motion of the agent towards the object but were in fact using some combination of the motion and linguistic contextual factors to attend to the object of interest.

A similar computation was done for the away trials, in order to determine whether participants initiated a saccade towards the target object before the agent left the scene. The difference between SOT and the point at which the agent exited the scene was computed. It was found that participants launched a saccade towards the target object 906 ms before the agent left the scene, indicating again that participants were not simply following the agent and deciding to look at the target object only once he or she had left. Here, however, there was no main effect of verb type (one-way repeated-measures ANOVAs), $F1(1, 29) = 2.35$, $p = .14$, $F2(1, 16) < 1$, $p = .84$. As also suggested by previous analyses, unless agent motion (toward) conforms with the expectations set by the linguistic utterance, verb effects do not emerge.

Target object saliency. These analyses examined the effects of target object saliency on eye movement behaviour. This information was based on the saliency ratings

obtained in a normative study used in the previous research (for a full description of how these norms were obtained, see Di Nardo, 2005). In short, however, participants were shown a still frame of each movie (one for each motion context, with the apparent direction of motion being unambiguously apparent) for 2 s and asked to list up to six objects they saw. The rating was computed by dividing the frequency with which the target object was listed by the total number of objects named, for each motion condition. The more frequently the target object was listed, the higher its visual salience within the scene was taken to be. Results showed that target object saliency ranged from 2.1% to 23.1%, i.e., this is how frequently participants listed the target object in the scenes after a 2 s exposition to each scene, suggesting that the target objects are not "popping out" in the scenes.

The first analysis explored the relation between target object saliency on the amount of time participants spent looking at target object. A Pearson's correlation was computed between target object saliency, the time spent fixating the target object before the verb-onset, after the verb-onset, and in total. Trials in which participants were already fixating the target object at verb-onset were excluded from this analysis because of the difficulty in clearly separating pre- and post-verbal fixations. We hypothesized that significant positive correlations would be obtained between the saliency ratings and the three fixation measurements, as objects that were more salient within the scene should draw more, or longer, fixations.

A one-tailed Pearson's correlation ($N = 402$) indicated that target object saliency ratings did correlate marginally significantly with the time spent looking at the target object after verb-onset ($r = .07$, $p = .09$) as well as the total amount of time ($r = .07$, $p =$

.08), but not with the time spent looking before verb-onset (r = .05, p = .14). This result partially supports our hypothesis; although more salient objects were fixated for longer durations, this occurred only after the disambiguating point. Given that the three motion conditions were equally represented among the two verb conditions, these results indicate that the linguistic constraints imposed by the latter half of the sentence drew more visual attention to the objects in question. Whether this was due to verb constraints or the actual utterance of the noun cannot be determined (note that Experiments 2 and 3 address this question more directly).

The second analysis explored the relation between target object saliency and the speed of post-verbal saccade initiation. We hypothesized that there would be a significant negative correlation, such that the more salient an object is rated to be, the more quickly the initial saccade should be launched. A one-tailed Pearson's correlation (N = 301) failed to confirm this hypothesis (r = - .06, p = .17). This indicates that even objects that stand out against a complex background do not attract fixations more quickly after the verb, suggesting that scene factors other than complexity take precedence in programming saccades. We presume these to be the dynamic elements event agent motion. Because this correlation was not significant, target object saliency was not included as a covariate in the main analysis described below.

The third analysis examined the relationship between target object saliency and whether or not the target object was being fixated at verb-onset. A point biserial correlation was computed between these two variables. A significant correlation was expected, because the more salient an object is, the more fixations it should attract at any given time, including at verb-onset. A one-tailed point biserial correlation (N = 458) was

computed, which indicated that there was no significant correlation (r = - .02, p = .66), contrary to our hypothesis. Again, it does not appear that more salient objects attract more fixations at the onset of the verb.

In sum, target object saliency only appears to affect eye movement behaviour when measured in terms of fixation time, and only in the post-verb segment of the trials. When measured in terms of SOT or probability of fixation at verb-onset, no effects of saliency were found.

Target event saliency. This set of analyses was similar to those described in the section above, except the target event saliency was employed in the place of the target object saliency. Target event saliency was computed in the normative study mentioned above, by asking half of the same set of participants what they thought was happening in the scene they just viewed, and half what they thought would happen next. As above, the basis for this rating was the frequency with which some variant of the target event (e.g., *The lady will break the eggs*) was listed. Results showed that on the whole, none of the target events was predictable, with frequency ratings ranging from 0 to 13.6%. This suggests that any effects of visual context on linguistic processing and any eye-movement directed by verb properties should be taken as effects of the unfolding linguistic and visual context in the dynamic scenes, and not on scene gist extracted from the configuration of objects and agents within the scenes.

The hypothesized results of these analyses were expected to be similar to those above: positive significant correlations for all three fixation duration correlations, because the more predictive a scene is of the target event structure, the longer the fixation duration on the target object. In addition, the higher the saliency rating, the more quickly

saccades should be initiated towards the target object after verb onset, and the more likely it should be fixated at any given time.

First, the relationship between target event saliency ratings and the three fixation durations was examined and was expected to yield significant positive correlations. This hypothesis was not supported: a Pearson's correlation ($N = 402$) indicated that the amount of time spent looking at the target object before verb-onset did not correlate significantly with target event saliency ($r = .05$, $p = .14$); nor did the time spent fixating after ($r = -.02$, $p = .66$), or the total time spent fixating the target ($r = -.02$, $p = .65$). Thus, only object saliency appears to have an effect on fixation times, which is likely to be driven by the interaction between linguistic and visual contexts.

Next, the relationship between target event saliency and SOT was computed. We hypothesized that there would be a significant negative correlation, which a Pearson's correlation ($N = 298$) did not confirm ($r = -.003$, $p = .48$). Again, more predictive events do not lead to faster SOTs. Because of this, target event saliency was not included as a covariate in the main analysis described below.

Finally, the relationship between target event saliency ratings and whether or not the target object was being fixated at verb-onset was examined. We expected that the more predictive a scene was in terms of the target event, the more likely the target object (implicated in the target event) would be fixated at verb-onset. A one-tailed point biserial correlation ($N = 458$) was computed, which indicated that there was no significant correlation ($r = .04$, $p = .18$), contrary to our hypothesis.

On the whole, these results suggest that scenes that are more predictive of the unfolding event utterance do not influence eye movement behaviour. While this can be

explained by the low target event saliency ratings, the more meaningful explanation suggests that the scenes used in this experiment contain such a high level of complexity that they preclude the anticipation of likely events. By implication, this also means that the objects specified by these events cannot be forecasted by scene composition alone.

Main analyses: Saccade onset time. These analyses examined the effect of verb type and motion type on the time taken to launch a saccade to the target object after the verb's utterance. The main analysis constituted a mixed factorial design, as all participants were exposed to all six of the experimental conditions (causative-away, causative-neutral, causative-towards, perception-away, perception-neutral, perception-towards), but only to one condition for each scene. Trials in which participants were already fixating the target object at verb-onset were not included in the analyses, nor were those in which participants never fixated the target object after verb-onset.

The main analysis examined the effect of verb type and motion type on SOTs.[1] This therefore constituted a 2 X 3 ANOVA. These analyses were conducted both by

[1] This analysis was to be conducted as an ANCOVA with target object saliency and target event saliency as the covariates, had these covariates correlated significantly negatively with the SOTs. This would be to eliminate any variability in the data due to differences in scene complexity and object/event salience. Because some objects could have been fixated more quickly than others based on these variables, the ANCOVA would control for any potential effects of scene complexity and object/event salience on the dependent variable, which may have led to an underestimation of the effects of the independent variables of interest (verb type and motion type). Given the lack of signification correlations, these covariates were not included.

participants (*F1*) and by items (*F2*), with any missing cell means replaced with the condition mean (note that in the participant analyses, a large number of such replacements were made, and therefore the item analysis should be considered more reliable wherever a discrepancy between *F1* and *F2* was found). We expected there to be a main effect of both verb type and motion type, but no interaction effect. More specifically, we expected SOTs to be faster in the causative condition than the perception condition, as well as in the towards condition than in the away and neutral conditions.

These hypotheses were only partially confirmed, as shown in Figure 4 (see Tables 3 and 4 for the ANOVA tables relevant to the analyses for this experiment). The results indicate that there was a significant main effect of motion type, $F1(2, 62) = 8.24$, $p = .0007$, $F2(2, 32) = 5.28$, $p = .01$, but not verb type, $F1(1, 31) < 1$, $p = .43$, $F2(1, 16) < 1$, $p = .38$. In addition, there was also a significant interaction effect in the participant analysis, $F1(2, 62) = 3.91$, $p = .02$, and a marginally significant interaction effect in the items analysis, $F2(2, 32) = 2.78$, $p = .08$. To follow up on the interaction between verb type and motion type, tests of the simple effects of verb type were conducted at each level of motion type. These showed that verb type only had a significant main effect in the toward condition, $F1(1, 31) = 5.30$, $p = .03$, $F2(1, 16) = 7.25$, $p = .02$, such that the causative group ($M1 = 956.1$, $SD1 = 477.5$, $M2 = 922.7$, $SD2 = 319.7$) led to shorter SOTs than the perception ($M1 = 1324.5$, $SD1 = 818.2$, $M2 = 1284.2$, $SD2 = 458.1$) group. No such effects were found in the away and neutral conditions.

This analysis provided only partial confirmation of our hypotheses. Although the effect of motion type was significant in the expected direction, with shorter SOTs in the toward condition, verb type did not produce robust effects. The only difference between

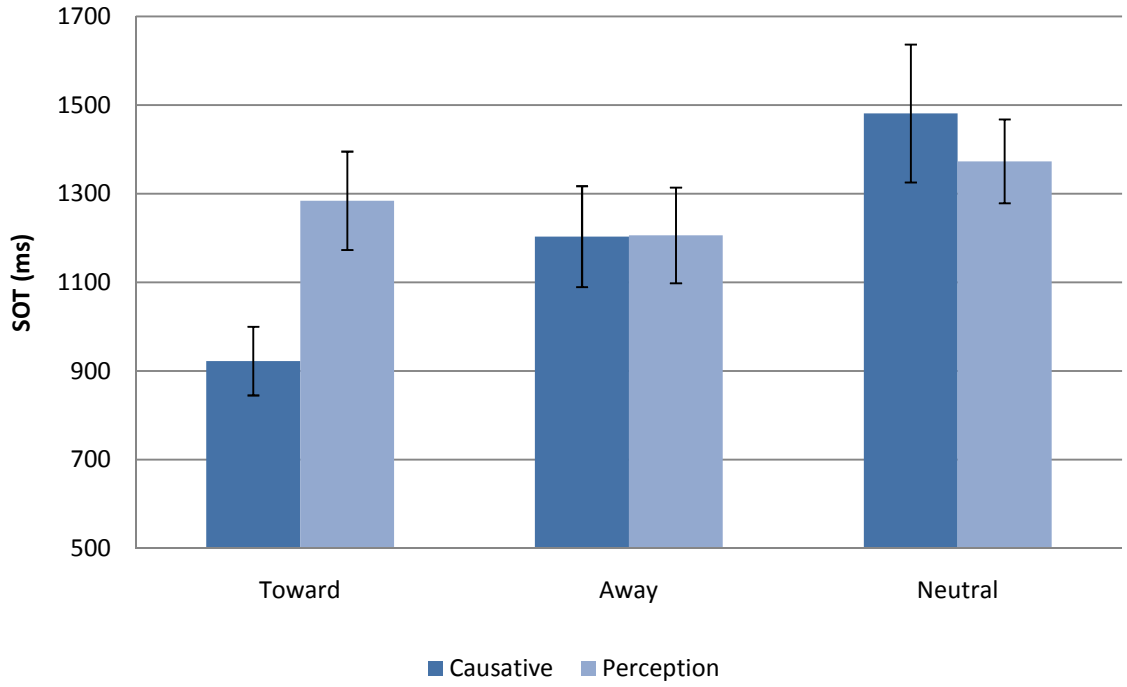


Figure 4. Effect of verb type and motion type on mean saccade onset times (SOTs) ± SE, computed by items. Verb type only had a significant main effect in the toward condition, $F2(1, 16) = 7.25$, $p = .02$.

causative and perception verbs was found in the toward condition. However, it is consistent with the results reported heretofore; wherever a verb effect was found, it has always been in the toward motion condition.

In addition to this main analysis, and to test more specifically for motion and verb effects, four planned comparisons were conducted (by participants, *t1*, and by items, *t2*). The first hypothesis was that the away and towards conditions would differ significantly, such that the toward condition would have a shorter mean SOT than the away condition. A one-tailed paired *t*-test was conducted to that effect, and revealed that there was a significant difference, $t1(31) = 1.88$, $p = .03$, $t2(16) = 1.79$, $p = .05$, as predicted. This is also consistent with the results found in the main analysis.

The second hypothesis was that the neutral and towards conditions would differ significantly, such that the toward condition would have a shorter mean SOT than the neutral condition. This hypothesis was confirmed with a one-tailed paired *t*-test, $t1(31) = 3.64$, $p = .0005$, $t2(16) = 3.51$, $p = .001$. Again, this is consistent the results reported above.

Third, to compare the two verb types without the confounding effects of the motion context, the causative-neutral and perception-neutral groups were compared. In the absence of any apparent motion in the scenes, we expected that the causative condition would lead to lower SOTs than the perception condition. A one-tailed paired *t*-test provided only limited support to this notion, $t1(31) = 1.47$, $p = .08$, $t2(16) = .56$, $p = .29$. Thus, in the absence of any agent motion, verb effects do not emerge, as was found in the analysis of early post-verb saccades.

Finally, to compare the two verb types in the toward condition, we compared the causative-towards and perception-towards conditions. We expected that the visual context (with the agent moving towards the target object) would aid in the semantic interpretation of the verb, such that SOTs would be lower in the causative-towards than in the perception-towards condition. A one-tailed paired *t*-test supported this notion, $t1(29) = -2.16$, $p = .02$, $t2(16) = -2.69$, $p = .008$, with the causative-towards condition having a significantly lower mean than the perception-towards condition. Again, this is consistent with several other results reported thus far, and suggests that the semantic consistency between the linguistic and motion contexts aids in the interpretation of the unfolding event.

The results of these main analyses suggest that in dynamic scenes, verb effects fail to emerge unless the motion context is consistent with the linguistic utterance (i.e., unless the agent moves toward the target object). Instead, the motion context seems to dominate the control of visual fixation, producing robust and consistent effects, with saccades being launched toward the target object when the human figure moves toward it rather than away, or doesn't move at all. These results are consistent with those of the previous analyses reported here, but fail to support the findings of Altmann and Kamide (1999) who did find verb effects in static ersatz scenes.

Comparison to Di Nardo (2005). To determine whether there was any effect of having replicated the earlier experiment with a larger visual angle on post-verb eye movement behaviour, two mixed-factor repeated -measures ANOVAs (by participants) were conducted. The factors for these analyses were 2 (experiment/visual angle; a between condition) x 2 (verb type; within) and 3 (motion type; within). In the first, we

hypothesized that visual angle would have a significant main effect, such that the SOTs would be slower in the larger visual angle condition (70.4°; current study) than in the smaller visual angle condition (50.4°; Di Nardo, 2005). As described earlier, this was because the larger visual angle of the scene should preclude any attentional shifts without corresponding eye movements. Results did not support this hypothesis; although visual angle had a significant main effect, $F(1, 68) = 4.57$, $p = .04$, it was in the direction opposite than that expected. SOTs were faster in the smaller visual angle condition (M = 1207.0, SD = 522.6) than in the larger visual angle condition (M = 1354.2, SD = 709.3). In other words, the increase in visual angle used in the current study seems to have caused participants to launch saccades to the target object later, not sooner. However, one possibility might be that the smaller visual angle in Di Nardo (2005) allowed for the parafoveal selection of the next fixation target (Findlay & Gilchrist, 2001), thus speeding up the saccade onset time following the disambiguating point. Given the larger visual angle of the present study's display, the covert scanning used by the visual attention system to search for a potential target would have been less likely to occur, and would result in longer saccade programming times in the search for the target object.

The second analysis examined the hypothesis that there would be fewer trials in which the target object was never fixated in the current study than in the previous study. If the visual system was attending to the target object without a concomitant fixation, and the increase in visual angle forced the production of such fixations, then more trials in which saccades were eventually launched should have resulted. Results again failed to confirm the hypothesis, with a significant main effect of visual angle, $F(1, 68) = 4.62$, $p = .03$ in the opposite direction from that predicted. There were in fact more trials in

which the target object was never fixated in the larger visual angle condition (M = .203, SD = .251) than in the smaller visual angle condition (M = .141, SD = .221). There are at least two possible explanations for these findings. First, as in the previous analysis, the greater visual angle may place the target object in peripheral vision, thus reducing the likelihood eye movements to that region. Second, because the actual size of the agent in the scenes is larger, and because human faces tend to attract attention (Morand, Grosbras, Caldara, & Harvey, 2010), participants may have remained fixated on the agent rather than shift their attention to the named object.

While these results were contrary to our expectations, they do suggest that attentional shifts were likely accompanied by saccades in the Di Nardo (2005) study. Therefore visual angle does not explain the lack of verb effects in both series of studies. Moreover, the fact that the body of findings from this study were essentially identical to those found by Di Nardo (2005) suggests that the visual-attentional system's insensitivity to the verb effect and the robust effect of the motion context are quite reliable with dynamic scenes. In fact, these results seem to indicate that visual fixation is primarily controlled by visual context, particularly the motion, and perhaps presence, of the agent. Given the relatively late saccade onsets, which as discussed previously fail to support the notion of anticipatory eye movements, one might conclude that they serve instead to *confirm* the interpretation of the unfolding utterance. The dynamic element present in the scenes thus seems to be the primary source of confirmatory evidence. However, given the unpredictability in the direction of motion (agents may move toward or away from the named object, or not at all), participants may in fact be adopting an approach in which they fail to fixate the target object until they receive confirmation from the agent's path

of motion. This is consistent with the findings of Knoeferle and Crocker (2007), who found that participants preferentially rely on depicted events over thematic knowledge encoded by the verb. The unpredictable path of the agent's motion likely introduced a cognitive set in which viewers relied heavily on agent motion to assist in the comprehension of the unfolding event.

Given the lack of supporting evidence for the verb effects found elsewhere in the literature, we can hypothesize that the use of dynamic scenes renders the visual world paradigm insensitive to verb-specific constraints. In addition, increasing the visual angle of the scenes does not enhance the sensitivity of the visual system to the effects of the verb. Nevertheless, the remainder of the experiments continued with this methodological adjustment, projecting the films onto the large screen, and explored whether other changes to the stimuli could produce a more robust verb effect. Specifically, we asked whether the presence of spoken language can at all influence the control of visual fixation to any degree. We also asked whether creating more semantic consistency between the event context and the sentence context would allow verb effects to emerge.

Experiment 2

Given the relative insensitivity to verb constraints found in our previous experiments in the dynamic visual world paradigm, the main purpose of this experiment was to determine whether the presence of spoken language has any effect at all on guiding eye movement behaviour. Specifically, we wanted to explore whether, and how, the absence of the linguistic context, i.e., the spoken sentences referring to the scenes' events, would alter the overall path of eye movements across the scenes. Despite the lack of verb effects found thus far, it is plausible to assume that the language context does

produce a scan path that is unique to situated language processing. At the very least, the utterance of the noun itself should direct visual attention to its visual referent, as this is the core assumption of the visual world paradigm. Indeed, several studies have found that eye movements tend to follow the utterance of any objects present within a visual array (e.g., Buswell, 1935; Yarbus, 1967; Tanenhaus et al.,1995; Altmann & Kamide, 1999).

To test this hypothesis, we compared eye movement behaviour across two conditions, one in which participants viewed the film clips, with the accompanying spoken sentences, as in Experiment 1 and in Di Nardo (2005, Experiment 2), and one without any spoken language at all. Specifically, we expected that in the absence of spoken language, saccade onset times would be longer than in the presence of language. We also expected fixation durations to the target object (both before and after the disambiguating point) to be longer in the presence of spoken language. Note that in the language-absent condition, the disambiguating point is considered to be the onset of motion of the agent, or its corresponding time point in the neutral condition. In this case, the neutral condition serves as a control condition both for motion and language contexts, allowing us to disentangle the effects of both of these variables.

In addition to this change to the linguistic context, the away condition was also excluded from this experiment. This was done for two reasons: one, to increase the statistical power of the experiment given the inclusion of a new independent variable (the presence or absence of the linguistic context); and two, because of its semantic inconsistency with the scene's event meaning. As discussed in Experiment 1, it is possible that the unpredictability of the agent's path of motion may create a cognitive set

in which participants rely heavily on this information to confirm their interpretation of the unfolding event. In eliminating the away condition, agents will be seen as either moving toward the target object or not at all. While the former condition is more consistent with the event meaning, the latter is not inconsistent, and thus serves as a control condition. Because the verbs used in the main clause are in the future tense, they refer to events yet to take place, and a lack of agent movement does not imply any inconsistency. Thus, viewers are less likely to be confused by the seemingly random movement of the agent, and to rely on this information exclusively to construct the event's meaning.

Distractor trials were introduced in this experiment to prevent participants from discerning the lack of variation in the experimental sentence structure (in the language-present condition). Should the underlying pattern (Patch Clause 1-NP1-*will*-Verb-NP2-RC-Patch Clause 2) have become apparent, it is possible that this might lead to a predictable path of viewing as they grew cognizant of the demands of the task over the course of the experiment. We therefore introduced several movies with a similar composition (realistic dynamic scenes, but without agent motion) that were accompanied by sentences of similar length but different structure. All other experimental variables were maintained constant (e.g., projection onto the large screen) to allow for valid comparison across studies.

Empirically, these changes should lead to an effect of verb type. In addition, if the language context is incrementally integrated with the motion context, we would expect faster SOTs in the language-present condition than in the language-absent condition, relative to the disambiguating point. We would also expect fixation durations

to the target object to be longer in the language-present condition, particularly after the disambiguating point when the noun's utterance takes place. Finally, we can directly contrast the *relative* contribution of the linguistic and visual contexts, which to our knowledge has never been attempted. This can be accomplished by comparing the time course of eye movements in the language-absent-neutral (control), language-absent-toward (language control), and language-present-neutral (motion control) conditions. If the visual context takes precedence over the linguistic context, then saccades should be faster in the motion control condition, whereas if the linguistic context takes precedence over the motion context, then saccades should be faster in the language control condition. Finally, if the visual and linguistic systems do interact in the construction of event meanings, then saccades should be fastest of all in the language-present-toward condition, which combines both effects.

Method

Participants. Thirty-eight participants took part in this study. They were all drawn from the Concordia University student body. None of these participants had taken part in earlier experiments employing similar materials. There were 35 females and three males, ranging in age from 18 to 34. Data from 34 of these participants were retained in the analyses, the rest having to be discarded due to computer calibration or recording errors. Inclusion and exclusion criteria were the same as in Experiment 1. All participants received course credit for their participation.

Materials and apparatus. The stimuli used in this experiment were identical to those used in Experiment 1. However, three notable changes were made to the experimental procedure. The first, and most critical to the purpose of this experiment,

was the introduction of a new experimental variable, language context. Fifteen of the participants were exposed to the spoken sentences, as in previous studies, while nineteen were not. Note that in the language-absent-neutral condition, there is no disambiguating point; no linguistic nor motion-related demarcation exists. However, for the purpose of this study, the disambiguating point was the same frame in the language-present and language-absent conditions, which was also the frame that corresponded to the verb onset in the language-present condition.

The second change involved the exclusion of the away condition. These trials were eliminated from the stimuli used in this experiment. There were thus a total of 68 sentence/movie pairs (2 verb types X 2 motion types X 17 scenes), distributed across 4 lists.

Finally, a series of six distractor trials were introduced that varied the sentence structure from that used in the experiment structure (patch clause-1, main clause, patch clause-2). Instead, sentences such as *The man is preparing an elaborate dinner to surprise his girlfriend for her birthday* or *Because he procrastinated, the student is staying up late writing a term paper that is due tomorrow* were used. Although some of these sentences still have a three-clause structure, they differed in the following respects: (1) the initial patch clause did not refer to some future event to take place within the scene; (2) there was often an absence of a target object, or one that had a visual referent within the scene; and (3) the final patch clause did not always make reference to the object of the verb in the previous clause. The sentences were recorded by a female English-speaking research assistant (different from the previous experiments) at a normal volume and pace using Audacity software (open source software) at a bitrate of 705 kbps.

The scenes used in these distractor trials were obtained from a series of studies conducted by van de Velde and her colleagues (van de Velde, 2008) using a similar visual world methodology. These were similar in content and composition to the experimental scenes; an agent was present in various everyday indoor and outdoor scenes (see Appendix D for a list of the scenes and sentences used in the distractor trials). However, the agent occupied the center of the screen, with two possible "target objects" on either side in the central plane. In addition, only "neutral" scenes were used in which there was no major change in position of the agent. Because these clips were recorded at a higher resolution (although presented at 720 x 480 pixels in NTSC format, with 29.97 frames per second), and had a different voice associated with the sentences (in the language-present condition), they did stand out from the rest. Although this could have caused participants to distinguish between the experimental and distractor trials, it also could have served the purpose of concealing the experimental conditions as well, keeping participants naïve to the purpose of the experiment.

As in Experiment 1, the film clips were projected onto the same blank screen using the same projector, at the same distance and dimensions. All other materials and apparatus were identical.

Procedure. The procedure used in Experiment 2 was similar to Experiment 1, except in one regard: participants in the language-absent condition did not wear the headphones during the experiment, and were thus not exposed to the sentences. Their instructions were slightly modified (see Appendix F) as they were told simply to watch the movies on the screen, and that their memory for these movies would be tested at the end of the experiment. They therefore did not expect to hear anything during the course

of the experiment. A short cued recall test was given at the end of the experiment to ensure participants paid attention during the trials. The cued recall task included only still frames from the scenes (presented or not), no sentences. The entire experiment lasted approximately 30 minutes.

Analyses. The same sets of analyses were used as in Experiment 1, with a few notable modifications. As in the previous study, we examined early eye movement behaviour, anticipatory eye movements, the effects of target object and target event saliency on eye movement behaviour, as well as the main analysis on SOT. However, in the present experiment, wherever ANOVAs were employed, the additional factor of language context was included (a between-subjects factor, creating the use of mixed-factor repeated-measures ANOVAs). In addition, wherever the effect of verb type was examined, we excluded the language-absent condition.

We also examined the effect of language context on the time spent looking at the target object, both before and after the disambiguating point, and in total. We expected that in the language-present condition, participants would spend more time looking at the target object, especially after the disambiguating point. Hypotheses for each set of analyses that include language context are presented below. Unless otherwise indicated, all hypotheses are the same as in Experiment 1.

Results and Discussion

All participants were retained in the analyses as the cued recall test scores ranged from 83.3% to 100%. As in Experiment 1, we also conducted a manipulation check to examine the effect of condition on the proportion of trials (by items) where the participant looked at the target object before the disambiguating point (the onset of the

verb in the trials with spoken sentences, the onset of motion in the towards condition, and the corresponding time point in the silent neutral condition). We did not expect a difference based on verb type or motion type because in these four conditions, as all sentences (in the language-present condition) were identical and all agents remained in the same position prior to the disambiguating point (see Figure 1). However, we did expect a main effect of language context; in the condition without spoken sentences, we expected the visual attention of participants to be less tied the actions of the agent, and hence result in a greater number of fixations to the target object early during scene processing.

As expected, neither verb type nor motion type had an effect on the proportion of trials in which participants launched a saccade to the target object prior to the disambiguating point ($p > .05$). However, as predicted, language context did have a significant main effect, $F(1, 32) = 8.00$, $p = .008$, such that there was a larger proportion of saccades to the target object prior to the disambiguating point in the language-absent condition than in the language-present condition. These results provide a first source of evidence that spoken language in the visual world paradigm does in fact alter the pattern of eye movements across dynamic scenes. In the absence of language, visual attention appears to be freed from its constraints and to foveate objects in the scene more unsystematically. In fact, in the absence of spoken sentences referring to future events, it may be that the role of the agent in effectuating those actions is less salient, and therefore features of the scene itself (and the objects that populate it) may have a greater influence on the control of eye fixation. As Henderson and Ferreira (2004) point out, scene knowledge (either short-term, gleaned from the composition of the present scene, or long-

term, obtained from previous, repeated exposure to similar scenes) serves as one of the sources of top-down influence on the determination of fixation location (the other major sources include information conveyed by the linguistic input). Thus the results of this analysis indicate that the targets of fixation are in fact determined by contribution of both the linguistic and visual contexts, or we would not have found a significant difference in the two conditions. The relative contribution of each source of information is addressed in the main analysis reported below.

Missing data. The proportions of missing data from the same three sources as in Experiment 1 were computed. As before, the first source of missing data was due to corrupt data, resulting from a system crash, poor calibration or, drift caused by head movements. Out of the 557 trials presented to participants, 173 (31.1%) were missing due to corrupt data. These trials were distributed evenly across the eight experimental conditions (no significant main effects or interactions were found).

The second source of missing data was trials in which participants never fixated the target object after verb-onset (as in Experiment 1, these saccades are not obligatory, and these trials are therefore not true "missing data" in the strictest sense of the word). One hundred and nineteen (21.4%) such trials were recorded. In order to examine whether language context, verb type and motion type had an effect on the proportion of trials (computed by participant; $N = 34$) where participants did not launch a saccade to the target object after verb-onset, a repeated-measures mixed factor 2 (language context) X 2 (verb type) X 2 (motion type) ANOVA was conducted. We expected that language context would have a main effect, such that there would be a higher proportion of trials where the target object was never fixated when language was absent than when it was

present. This should be especially true in the neutral condition, where the absence of the linguistic and motion contexts should produce a higher proportion of such trials.

Results indicated that there was a marginally significant interaction effect between the three independent variables, $F(1, 32) = 3.75$, $p = .07$. To interpret this result, two 2 (verb type) X 2 (motion type) ANOVAs were conducted at each level of language context. The first indicated that motion type had a significant main effect in the language-absent condition, $F(1, 18) = 34.70$, $p < .0001$, such that there was a significantly higher proportion of "never looked" trials in the neutral condition than the toward condition. Thus, in the absence of auditory or visual cues to direct visual attention toward the target object, participants were less likely to fixate the object referent, which confirms our hypothesis.

The second ANOVA showed that the interaction between verb type and motion type was marginally significant in the language-present condition, $F(1, 14) = 3.72$, $p = .07$. A follow-up one-way repeated-measures ANOVA indicated that in the language-present-neutral condition, verb type did have a significant main effect, $F(1, 14) = 6.14$, $p = .03$, such that participants were more likely to fixate the target object after the verb-onset in the causative condition ($M = .131$, $SD = .207$) than in the perception condition ($M = .295$, $SD = .238$). However, verb type did not have an effect in the language-present-toward condition ($p > .05$). These results indicate when spoken language was present, but motion absent, hearing the causative verb lead to more fixations to the target object than the perception verb. However, no verb effect was found when motion was present.

The results of this analysis show that after the disambiguating point, the target object is less likely to be fixated when language is absent and there is no movement of the agent. Taken together with the results of the previous analysis, which showed that the target object is *more* likely to be fixated before the disambiguating point when language is absent, it appears that once fixated, this object is not used in the construction of event meaning when there is an absence of visual or linguistic cues to suggest its involvement. Again, this serves as evidence that the presence of language does in fact alter the pattern of eye movements across a given scene. As to the finding that in the presence of language, but not motion, causative verbs led to more fixations to the target object, this does not support the findings reported in Experiment 1. No such effect was found in the equivalent analysis, and wherever verb effects were found, they tended to be prompted by the toward motion context; however, the reduced power of this analysis may account for these findings. Nevertheless, they do point to the limited evidence that verb effects can emerge under some conditions.

A Chi-Square test was conducted to examine the effect of whether having looked at the target object before the disambiguating point might have led to fewer fixations after this point. As in Experiment 1, there was no significant effect of fixations to the target object prior to the disambiguating point on the number of those occurring after this point; $\chi = 1.82$, $p = .18$. Thus, failure to fixate the target object after the disambiguating point is not due to having perceived earlier.

The third source of missing data derived from trials in which participants happened to be fixating the target object at verb-onset. This occurred in 31 (5.6%) of the trials. These trials had to be excluded from any analyses examining the effect of verb type

or motion type on subsequent eye movement behaviour, because of the inability to discern whether participants continued to fixate the object because they were cued by the verb or motion context or not. A 2 (language context) X 2 (verb type) X 2 (motion type) mixed-factor repeated-measures ANOVA ($N = 34$) was conducted to examine the effect of these factors on the proportion of these trials (by participants). We did not expect to find any significant main effects of verb type (in the language-present condition) or motion type, as prior to the disambiguating point, these four conditions were identical. However, to the extent that the initial segment of the sentence creates an event representation including the target object, we might expect language context to have a main effect, such that it would be more likely for the target object to be fixated at verb-onset in the language-present condition.

The results indicated that neither language context nor verb type had a main effect ($p > .05$), although motion type did, $F(1, 32) = 4.18$, $p = .05$. These results confirm the idea that the verb class of the sentence did not affect whether participants were fixating the target object at the time of the disambiguating point. However, it does appear that there was a significantly higher proportion of trials in which participants were already looking at the object at the disambiguating point in the neutral condition than in the toward condition, contrary to our expectation. This is difficult to explain given that the movie clips were essentially identical up to that point, although the results might have been spurious. In addition, the hypothesis that language context would have an effect was not confirmed. Thus, it appears that the initial patch clause does not, in general, provide any information that might be predictive of the target object.

In summary, the pattern of missing data attributable to two of the three sources (trials where participants never looked or were fixating the target object at verb-onset) were not randomly distributed across the eight conditions. Because such a large proportion of data was missing, the analyses reported below (both by participants [*F1*] and by items [*F2*]) had any missing cell means replaced with the condition mean.

Effect of language context on fixations to target objects. In order to test the hypothesis that the presence of language would lead to longer fixation times to the target object, particularly after the disambiguating point, we correlated language context and the time spent fixating the target object before the disambiguating point, the time after the disambiguating point and the total time spent fixating. We expected that there would be a significant positive correlation, such that in the language-present condition, the target object would be fixated for longer intervals. A one-tailed point biserial correlation (N = 353) was computed, which indicated that there was no significant correlation between the three time measurements and language context (before: r = - .07, p = .91; after: r = - .04, p = .77; total: r = - .06, p = .86). These results suggest that the presence of spoken language, including utterances that specifically name an object within a scene, do not influence fixation durations to that object. We had hypothesized in the introduction to this experiment that situated spoken language should influence the pattern of eye movements across the scene on two fronts: fixation durations, which this analysis failed to confirm, and the speed of saccade initiation. The next two sets of analyses address this question more directly.

Analysis of early post-verb cumulative saccades to the target object. The effects of language context, verb type, motion type and sentence point on the cumulative

proportion of saccades to target object after verb-onset were examined using a mixed-factor repeated-measures 2 (language context: present *vs.* absent; between-subjects factor) X 3 (sentence point: verb-offset, noun-onset, and noun-offset) X 2 (verb type: causative *vs.* perception) X 2 (motion type: neutral *vs.* towards) ANOVA (both by participants, *F1*, and by items, *F2*). We hypothesized that there would be significant main effects for sentence point, language context and motion context, or a significant interaction between two or more of these variables, because the cumulative number of saccades should increase as the sentence unfolds, particularly in the toward and causative conditions. Note that in the language-absent-neutral condition, which does not possess a perceptually distinguishable disambiguating point, saccades were counted from the same frame used in the other films (standardized across all eight conditions). Because the effects of verb type are confounded in the language-absent condition, we did not interpret any significant effects in the results of the overall ANOVA.

The results (which are plotted in Figures 5 and 6) partially confirmed our hypotheses; the main effects of sentence point, $F1(2, 64) = 36.82$, $p < .0001$, $F2(2, 32) = 45.95$, $p < .0001$, and motion type, $F1(1, 32) = 6.53$, $p = .02$, $F2(1, 32) = 5.75$, $p = .02$, were significant, while the main effect of verb type was not, $F1(1, 32) = 3.36$, $p = .08$, $F2(1, 32) < 1$, $p = .19$. In addition, the main effect of language context was also significant, $F1(1, 32) = 7.03$, $p = .01$, $F2(1, 32) = 12.81$, $p = .001$, such that the language-present condition ($M1 = .174$, $SD1 = .274$, $M2 = .184$, $SD2 = .315$) had a higher cumulative proportion of saccades than the language-absent condition ($M1 = .086$, $SD1 = .210$, $M2 = .070$, $SD2 = .175$), as expected. Finally, the three-way interaction effects between sentence point, verb type and motion type was significant (see Table 5 and 6 for

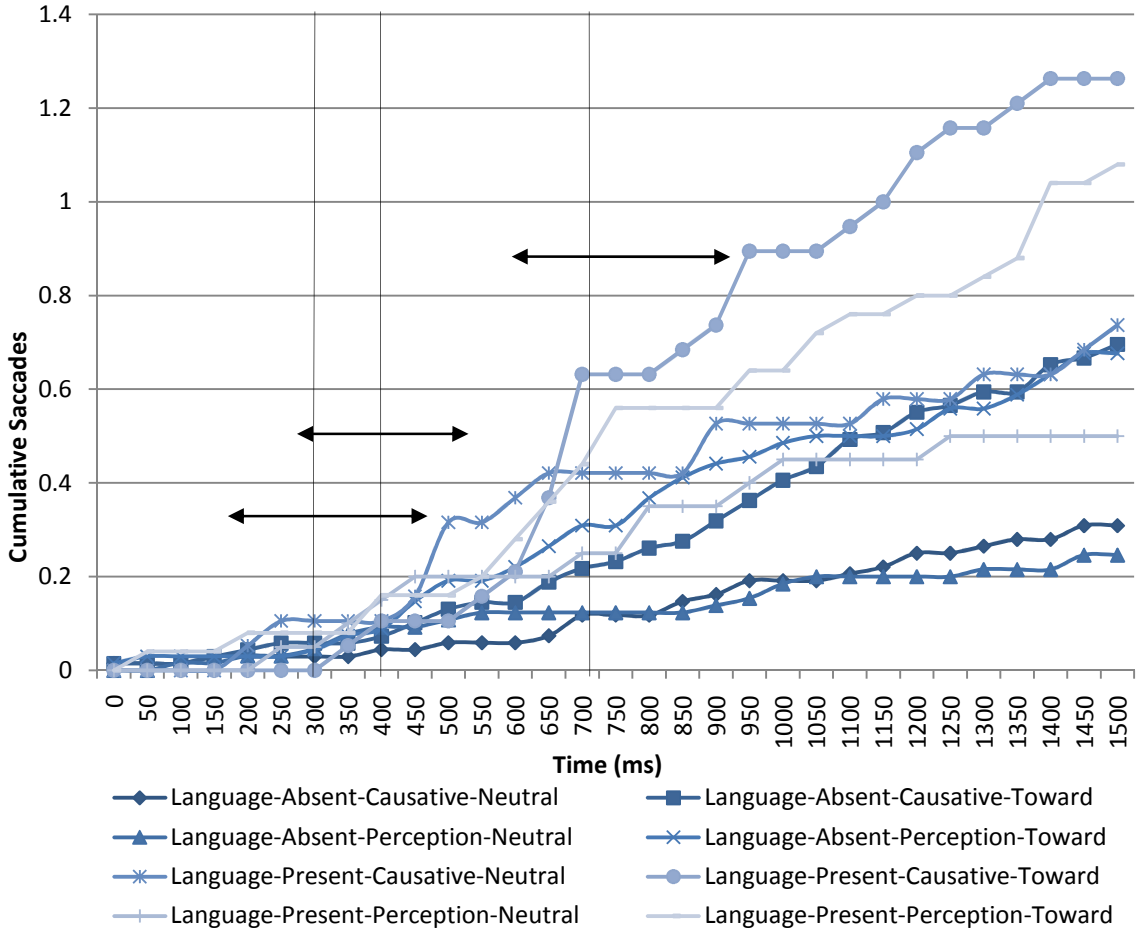


Figure 5. A plot of the mean cumulative average number of fixations (by participant) to the target object after verb-onset. Each line refers to a single condition, and each point to one 50-ms bin. The origin of the X-axis refers to the verb-onset, and the three vertical lines mark the temporal boundaries of the verb and noun (average onset and offset). The double-headed horizontal arrows on each boundary indicate the range of onsets and offsets at the points in time relative to the verb-onset, taking into account the variable lengths of each of the sentence segments (i.e. verbs and noun phrases). The point at which each of the coloured lines (referring to cumulative fixations for each condition) intersects with the three critical sentence points (verb-offset, noun-onset and noun-offset) were computed and compared in an ANOVA.

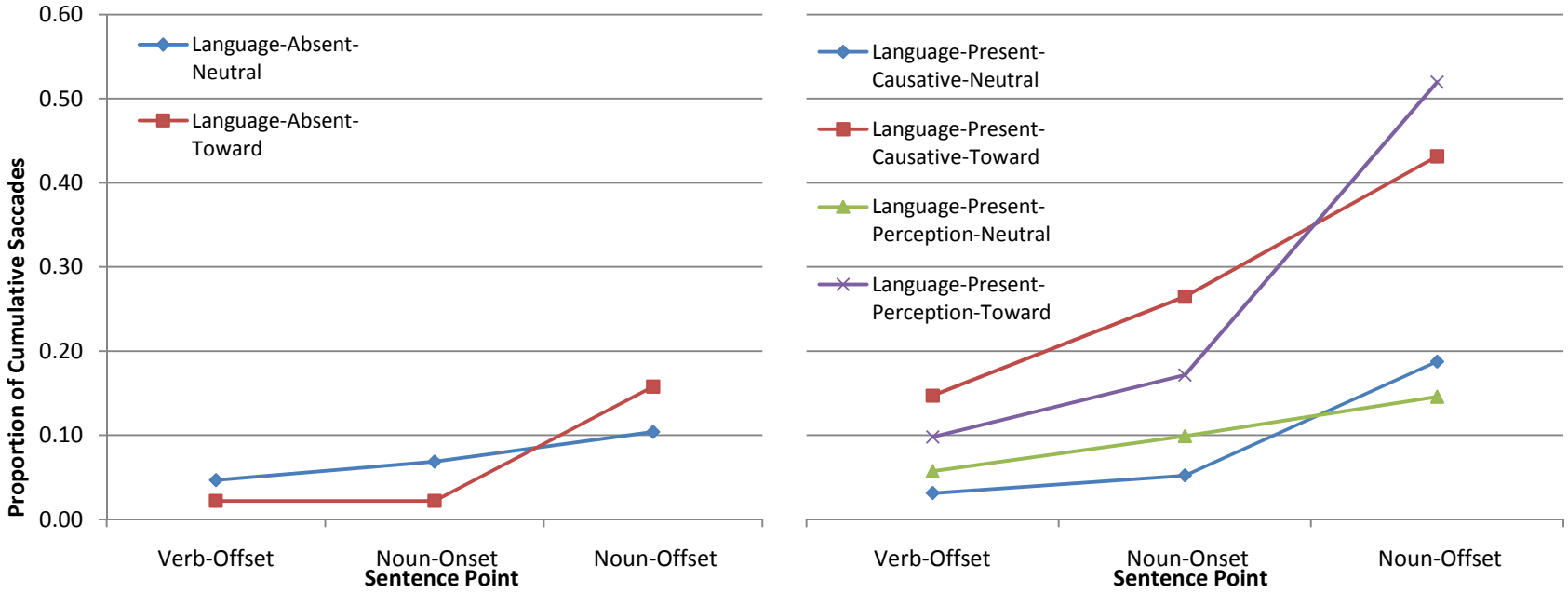


Figure 6. Mean number of cumulative saccades towards target object at each of the three critical sentence points, verb-offset, noun-onset and noun-offset (computed by items), for each level of language context (language-absent and language-present). A cursory examination of the graph shows that the two language-absent conditions (collapsed across verb type) and the two language-present-neutral conditions fail to reach the same magnitude as the two language-present-towards conditions, which exhibit very early post-verb effects.

the ANOVA tables relevant to this analysis), $F1(2, 64) = 3.74$, $p = .04$, $F2(2, 64) = 3.72$, $p = .03$.

In order to explore the three-way interaction between sentence point, verb type and motion type, tests of the simple effects of language context and motion type (but not verb type, due to the absence of language in one of the two language context conditions) across all levels of sentence point were conducted. The first of these tests indicated that there was a marginally significant interaction between language context and motion type at the verb offset, but only in the item analysis, $F1(1, 32) < 1$, $p = .59$, $F2(1, 32) = 2.96$, $p = .09$. Further tests failed to show any significant effects of motion type at both levels of language context at the verb offset. These results indicate that the effects of language context and motion type fail to have any robust effect by the offset of the verb, but before the noun was uttered.

At noun-onset, again, language context and motion type had a significant interaction effect in the item analysis, $F2(1, 32) = 5.99$, $p = .02$, and a marginally significant interaction effect in the participant analysis, $F1(1, 32) = 3.47$, $p = .07$. Further tests indicated that motion type only had a marginally significant main effect in the language-present condition at noun-onset in the item analysis, $F2(1, 16) = 4.24$, $p = .06$, but not in the participant analysis, $p = .14$. The effect was not significant in the language-absent analysis, $p > .10$. Taken together, these results indicate that in the presence of spoken language, by the beginning of the utterance of the noun, the onset of motion of the agent has begun to bias visual attention toward the target object. However, this result was not consistent in the two analyses by participants and items.

Finally, at the offset of the noun, language context and motion type again had a significant interaction effect, $F1(1, 32) = 2.76$, $p = .06$, $F2(1, 32) = 5.45$, $p = .02$. As in the previous test, motion type only had a significant main effect in the language-present condition at noun-offset, $F1(1, 16) = 11.30$, $p = .005$, $F2(1, 16) = 15.29$, $p = .001$. Again, this indicates that by the offset of the noun in the spoken language condition, the agent's motion toward the target object does trigger eye movements to that object.

Taken together, the results of this ANOVA show two main points. The first is that between the onset of the verb and offset of the noun, the presence of spoken language has already begun to show very early effects on the number of saccades launched to the target object, clearly biasing the visual-attentional system toward that object, even as its noun referent is being uttered. The second point is that in the presence of that utterance, the onset of motion in the agent also triggers saccades towards the target object, but not in the absence of spoken language. These results clearly show that the language context does indeed influence eye movement behaviour at the very earliest stages of event processing, and that in fact, at this early stage, the motion context fails to have any influence without the co-present linguistic context. This can be taken as evidence for incremental integration between these two systems, as evidenced by the visual-attentional system. Furthermore, it gives some limited evidence for the primacy of the linguistic context over the motion context in guiding eye movements, in the brief moments following the disambiguating point.

Next, to further examine the effects of verb type in the language-present condition, a 3 (sentence point) X 2 (verb type) X 2 (motion type) repeated-measures ANOVA was conducted. We hypothesized that there would be a significant three-way

interaction effect between sentence point, verb type and motion type in the language-present condition, such that the mean number of cumulative saccades to the target object would increase more at each sentence point for the causative condition than the perception condition, and for the toward condition than the neutral condition. This is because the interpretation of the verb and the noun phrase should proceed incrementally, thus providing more restrictive information as time proceeds, especially in the more semantically restrictive causative and toward conditions. This would indicate that verb-specific information can constrain visual attention at the very early stages of processing, as can the agent's early movement toward the target object.

Results indicated that again, the main effects of sentence point, $F1(2, 28) = 23.46$, $p < .0001$, $F2(2, 32) = 37.22$, $p < .0001$, and motion type, $F1(1, 14) = 8.79$, $p = .01$, $F2(1, 16) = 9.85$, $p = .006$, were significant, although the main effect of verb type was not, $F1(1, 14) < 1$, $p = .34$, $F2(1, 16) < 1$, $p = .96$. In addition, the three-way interaction was again significant, but only in the item analysis, $F2(2, 32) = 3.94$, $p = .03$. In the participant analysis, the three-way interaction was not significant, $F1(2, 28) = 1.11$, $p = .34$, but the interaction between sentence point and motion type was, $F1(2, 28) = 7.77$, $p = .002$ (the main effect of verb type was not significant, $F1(1, 14) < 1$, $p = .34$, contrary to our hypothesis). To explore the effects of verb type and motion type at each level of sentence point, a series of three 2 (verb type) X 2 (motion type) repeated-measures ANOVAs were conducted. The first indicated that neither verb type nor motion type had a significant main effect at verb-offset, $p > .05$. However, motion type did have a marginally significant main effect at noun-onset in the analysis by items, $F2(1, 16) = 4.24$, $p = .06$, but not in the analysis by participants, $F1(1, 14) = 2.41$, $p = .14$. At noun-

offset, motion type had a significant main effect in the analyses by items and participants, $F1(1, 14) = 11.23$, $p = .005$, $F2(1, 16) = 15.29$, $p = .001$, Verb type did not have a significant main effect in any of these ANOVAs (nor were the interactions significant, $p > .05$).

These results show that during the utterance of the noun, the onset of motion in the agent does trigger saccades toward the target object, as in the analysis reported previous to this one. However, it does not appear that verb type shows any effect at this early phase of sentence processing, at least as measured by eye movement behaviour, even in the presence of agent motion.

Planned comparisons were also conducted to test a series of hypotheses regarding these early post-disambiguating point saccades. These consisted of a series of three paired *t*-test computed both by participants (*t1*) and by items (*t2*). First, to determine whether motion context has an effect soon after the disambiguating point, we compared the toward and neutral conditions across all sentence points, language contexts and verb types. We expected that the mean number of cumulative saccades would be higher in the toward condition than in the neutral condition. We used a one-tailed paired *t*-test to compare these two conditions. Our hypothesis was supported: the toward condition ($M1 = .164$, $SD1 = .169$, $M2 = .170$, $SD2 = .326$) had significantly more fixations to the target object than the neutral condition ($M1 = .103$, $SD1 = .154$, $M2 = .084$, $SD2 = .164$), $t1(33) = -1.83$, $p = .04$, $t2(16) = -2.01$, $p = .03$. This suggests that during the period following the disambiguating point, both when language is present and when it is absent, the onset of agent motion can direct eye movements toward the appropriate object referent, a finding consistent with the results of the ANOVA reported above.

To examine the difference between causative and perception verbs without the moderating effect of motion, we compared the causative-neutral and perception-neutral groups across all sentence points in the language-present condition. We hypothesized that the causative-neutral would have a significantly higher mean than the perception-neutral group. A one-tailed paired *t*-test failed to show a significant difference, $t1(14) = -.17$, $p = .56$, $t2(16) = .76$, $p = .22$. This suggests that verb-specific constraints in the absence of agent motion do not influence eye movement behaviour at this early stage of processing after the disambiguating point, which replicates the finding of the equivalent analysis in Experiment 1.

Finally, we compared the causative-towards and perception-towards groups across all sentence points in the language-present condition, with the expectation that the mean would be higher in the former than the latter. This hypothesis was also not supported, $t1(14) = -.96$, $p = .82$, $t2(16) = -.82$, $p = .79$. These results indicate that even in the presence of agent motion, verb-specific information fails to influence eye movement behaviour by the end of the noun's utterance, which does not support the results from Experiment 1, although lower power may account for this finding.

Taken together, these results suggest that in the presence of language, only agent motion can guide visual attention toward the target object. Verb constraints, whether with or without corroborating evidence from the motion context, have no effect. These results are on the whole similar to those of the first experiment, but contrary to the findings reported by Altmann and Kamide (1999). While the motion context does bias focal attention toward the target object, the results reported here do not indicate whether

these eye movements are considered evidence for anticipatory effects. The series of results reported below will help to illuminate this issue more clearly.

Anticipatory eye movements. In order to determine whether any anticipatory eye movements may have been made, saccade onset times (SOTs) were compared to two other time points in the sentences: the noun-onset and the noun-offset. In addition to significant main effects of both verb and motion type, we expected language context to also have a significant main effect, such that shorter SOTs would be produced in the language-present condition. In addition, we calculated the proportion of trials in which a saccade was launched towards the target object before the onset and at the offset of the noun. In the language-absent condition, the analyses used calculations from the time points equivalent to the onsets and offsets of the noun, to make for valid comparisons across the two language context conditions.

We did not find any evidence for anticipatory eye movements. On average, eye movements were initiated 942 ms after the noun-onset in the causative condition, and 748 ms after the noun-onset in the perception condition. Relative to the noun-offset, saccades were launched 661 ms after this linguistic boundary in the causative condition, and 468 ms after this point in the perception condition. In addition, saccades were launched toward the target object before the onset of the noun in 6.6% of the causative trials and 9.1% of the perception trials. However, saccades were launched toward the target object before the offset of the noun in 11.6% of the causative trials and 20.7% of the perception trials. A one-tailed paired *t*-test showed that the perception condition had a significantly higher proportion of such anticipatory saccades than the causative condition at noun-onset, $t(16) = -1.71$, $p = .05$, as at noun-offset, $t(16) = -1.66$, $p = .06$ (although this

difference was marginal). This contrary to what might be expected, as the less semantically restrictive perception verb does not possess any information that might promote such anticipatory eye movements. In addition, it is in contrast to the findings of Experiment 1, where no difference was found. Thus, given the fact that the effect was only marginally significant, it may represent an instance of Type II error.

To further explore the time course of these eye movements, we examined the difference in time between the onset of the noun and SOT. This was subjected to a 2 (language context) X 2 (verb type) X 2 (motion type) repeated-measures ANOVA, in order to determine whether these first saccades were affected by these three variables. Results indicated that the main effect of language context was significant, $F1(1, 32) = 13.32$, $p = .0009$, $F2(1, 32) = 7.49$, $p = .01$, such that saccades were launched more quickly in the language-present condition ($M1 = 640.6$, $SD1 = 490.1$, $M2 = 646.7$, $SD2 = 596.0$) than the language-absent condition ($M1 = 973.9$, $SD1 = 598.0$, $M2 = 1020.1$, $SD2 = 670.0$). The main effect of verb type was also significant in the participant analysis, $F1(1, 32) = 4.90$, $p = .03$, but not the item analysis, $F1(1, 32) = 1.38$, $p = .25$. The main effect of motion type was not significant, $F1(1, 32) = 2.73$, $p = .11$, $F2(1, 32) = 3.08$, $p = .09$. In addition, the three-way interaction between language context, motion type and verb type was significant in the participant analysis, $F1(1, 32) = 4.90$, $p = .03$, but not in the item analysis, $F2(1, 32) < 1$, $p = .59$. However, the interaction between language context and motion type was significant in the item analysis, $F2(1, 32) = 6.25$, $p = .02$.

To further explore the significant interactions, tests of the simple effects of verb type (but only in the language-present condition) at each level of the language context and motion type were conducted. These showed that the main effect of motion type was

not significant in the language-absent condition, $F1(1, 18) < 1$, $p = .81$, $F2(1, 16) < 1$, $p = .62$. In the language-present condition, motion type did have a significant main effect, $F1(1, 14) = 8.76$, $p = .01$, $F2(1, 16) = 10.03$, $p = .006$, although verb type did not ($p > .05$).

This analysis was repeated for the difference in time between the offset of the noun and SOT. A similar pattern of results was found: again, language context had a significant main effect, $F1(1, 32) = 12.40$, $p = .001$, $F1(1, 32) = 7.97$, $p = .008$, such that the language-present group ($M1 = 353.8$, $SD1 = 481.4$, $M2 = 360.7$, $SD2 = 591.2$) had shorter SOTs than the language absent group ($M1 = 694.3$, $SD1 = 639.9$, $M2 = 750.4$, $SD2 = 694.9$). In addition, the main effect of verb type was significant in the participant analysis, $F1(1, 32) = 6.62$, $p = .01$, but not the item analysis, $F2(1, 32) = 2.22$, $p = .15$. The main effect of motion type was also not significant, $F1(1, 32) = 2.81$, $p = .10$, $F2(1, 32) = 3.18$, $p = .08$. Again, the interaction between all three factors was significant in the participant analysis, $F1(1, 32) = 5.78$, $p = .02$, while the interaction between motion type and language context was significant in the item analysis, $F2(1, 32) = 5.03$, $p = .03$. Further analyses revealed that the main effect of motion type was significant only in the language-present condition, $F1(1, 14) = 7.34$, $p = .02$, $F2(1, 16) = 9.32$, $p = .008$.

These results indicate that at both the onset and offset of the noun, motion context does have an effect on how quickly the target object is fixated, but only in the condition in which language is present, which is consistent with the findings of Experiment 1. When language is absent, motion type does not influence the speed of eye movements. Again, this suggests that linguistic and visual information are incrementally integrated as the event unfolds, and that the visual system fails to take advantage of the information

conveyed by the motion context when the spoken sentence is not co-present. However, it appears that verb-specific effects do not emerge in the language-present context, as in Experiment 1. Thus, it appears that while the visual attention system preferentially relies on the motion context only when language is present, the restrictions imposed by verb class are not sufficient to further elicit faster eye movements.

In order to determine whether participants were able to anticipate the object that the sentence referred to by initiating a saccade towards it before the agent in the scene reached it (again, in the towards condition only), the difference between SOT and the time at which the agent touched the object was computed. It was found that participants launched a saccade towards the target object 2272 ms before the agent made contact with the object, indicating that participants were using some combination of the motion and linguistic contexts to anticipate the target object rather than simply following the motion of the agent towards the object.

To determine whether the effects of language context and verb type affect the time course of these eye movements, a 2 X 2 mixed-factor repeated-measures ANOVA was conducted. We hypothesized that language context and verb type would show a significant interaction effect, such that the reaction times would be shorter in the language-present condition than the language-absent condition, as the presence of spoken language should bias the visual-attentional system toward the event structure's patient role filler. We expected the same to be true for the verb type, such that causative verbs should also lead to faster saccades in the language-present condition than the perception condition. Results did not reveal a significant interaction between language context and verb type, nor a main effect of verb type ($p > .05$), contrary to our hypotheses. However,

there was a significant main effect of language context, $F1(1, 32) = 45.97$, $p < .0001$, $F2(1, 32) = 11.52$, $p = .002$, such that eye movements were launched more quickly in the language-present condition ($M1 = -1141.3$, $SD1 = 485.1$, $M2 = -1177.7$, $SD2 = 772.7$) than in the language-absent condition ($M1 = -395.8$, $SD1 = 465.\ 7$, $M2 = -458.0$, $SD2 = 626.8$). Thus, when the agent was moving toward the target object, the presence of spoken language caused participants to initiate saccades toward that object much more quickly than when language was not present. This indicates that the event representation constructed by the linguistic utterance is integrated with the agent's path of motion to influence visual attention, but is not influenced by representations encoded by the verb, similar to the findings of Experiment 1. While these eye movements did arrive at the target object well before the agent did, they are not true anticipatory eye movements in the sense that they occurred after the noun referent was uttered. However, they can be seen as an attempt to confirm the visually unfolding event referred to by the spoken sentence.

Target object saliency. The first analysis examined the correlation between target object saliency ratings, pre-verb fixation durations, post-verb fixation durations and total fixation durations. We hypothesized that the saliency ratings would correlate positively and significantly with the amount of time spent looking at the target objects, as these ratings should draw more, and/or longer, fixations. A one-tailed Pearson's correlation ($N = 353$) indicated that target object saliency ratings did not correlate significantly with the time spent looking at the target object before verb-onset ($r = -.01$, $p = .57$), nor the time spent looking after verb-onset ($r = .05$, $p = .17$), or the total amount of time ($r = .05$, $p = .19$). These results do not support our hypotheses.

As this same analysis in Experiment 1 found a positive correlation with fixation durations after the disambiguating point only, prompting the question as to whether it is the utterance of the noun that led to this relationship, we re-ran these correlations separately for each language context. A one-tailed Pearson's correlation showed that in the language-absent condition (N = 200), target object saliency did not correlate significantly with the time spent fixating after the disambiguating point (r = .06, p = .18), nor was the correlation significant in the language-present condition (N = 153), r = .04, p = .32. This may be due to the smaller sample size, as well the elimination of the away motion condition or the introduction of the fillers.

In the next set of analyses, the relationship between target object saliency and SOT was computed. We hypothesized that there would be a significant negative correlation, with saccades being launched more quickly when the target object is more salient. A one-tailed Pearson's correlation (N = 234) did not support this hypothesis (r = - .08, p = .12). As in Experiment 1, it appears that more salient object do not attract fixations more quickly after the disambiguating point. However, we again conducted this correlation separately for each language context to determine whether the pattern would differ across both conditions. A one-tailed Pearson's correlation (N = 113) showed that in the language-absent condition, target object saliency did not correlate significantly with SOT (r = - .06, p = .25). In the language-present condition (N = 121), this correlation also proved to be non-significant (r = - .08, p = .18). Because this correlation was not significant, target object saliency was not included as a covariate in the main analysis described below.

Third, the relationship between target object saliency ratings and whether or not the target object was being fixated at verb-onset was examined. We expected that there would be a significant positive correlation, such that the higher the saliency rating, the more likely the target object would be fixated at verb-onset. A one-tailed point biserial correlation ($N = 384$) was computed, which indicated that there was no significant correlation ($r = -.03$, $p = .73$), contrary to our hypothesis.

Target event saliency. The same set of analyses described above was conducted with target event saliency instead of target object saliency. First, the relationship between target event saliency ratings and the three fixation durations was examined and was expected to reveal significant positive correlations, as the more predictive a scene is of the event about to unfold, the longer the object that is the target of that event should be fixated. A Pearson's correlation ($N = 353$) indicated that target event saliency did not correlate significantly with any of the three fixation durations; the amount of time spent looking at the target object before verb-onset ($r = -.02$, $p = .63$), the time spent fixating after ($r = -.01$, $p = .60$), and the total time spent fixating ($r = -.02$, $p = .63$) failed to show a significant relationship with event saliency. These results are similar to Experiment 1, which also did not reveal significant positive correlations, and indicate that the target event saliency was not sufficient to influence fixation durations.

Next, we computed the relationship between target event saliency and SOT, hypothesizing that there would be a significant negative correlation, such that more predictable events would trigger faster saccades to the target object. A Pearson's correlation ($N = 234$) did not confirm this hypothesis ($r = -.06$, $p = .17$). This replicated the result found in Experiment 1, and again, target event saliency was not included as a

covariate in the main analysis. Because the events are not predictable, it can be said that target selection is a function of linguistic and motion effects instead.

Finally, we tested the hypothesis that target event saliency ratings would determine whether or not the target object was being fixated at the disambiguating point. We hypothesized that the more a scene was predictive of the target event, the more likely the target object would be fixated at the disambiguating. A one-tailed point biserial correlation ($N = 384$) did not show a significant correlation ($r = -.02$, $p = .67$), contrary to our hypothesis, but consistent with Experiment 1.

Overall, these results replicated our findings from Experiment 1. Target event saliency does not correlate significantly with the three main manifestations of eye movement behaviour; namely, fixation durations, saccade onset times, and probability of fixation at the disambiguating point. We take these results to further strengthen the notion that the scenes do not contain enough information to predictably generate a consistent event meaning, particularly one that involves the target object.

Main analyses: Saccade onset time. Again, we examined the effect of verb type and motion type on saccade onset times (SOTs), in addition to the new factor of language context (both by participants, *F1*, and by items, *F2*, with any missing cell means replaced by the condition mean). We hypothesized that language context would have a significant main effect on SOT, with the language-present condition producing shorter SOTs than the language-absent condition. We also expected motion type to have a significant main effect on SOT, with faster saccades in the toward condition than in the neutral condition. We did not test the effects of verb type, as this variable was confounded in the language-

absent condition. A separate analysis with just the data from the language-present condition was conducted to examine the effects of verb type.

A 2 (language context) X 2 (verb type) X 2 (motion type) mixed-factor repeated-measures ANOVA showed that our hypothesis was partially confirmed, as shown in Figure 7 (see Table 7 and 8 for the ANOVA tables relevant to the analyses for this experiment). The results indicated that language context, $F1(1, 32) = 14.38$, $p = .0006$, $F2(1, 32) = 7.56$, $p = .01$, had a significant main effect, while verb type did not, $F1(1, 32) = 1.55$, $p = .22$, $F2(1, 32) < 1$, $p = .64$. There was a trend toward significance for the main effect of motion type, $F1(1, 32) = 2.62$, $p = .11$, $F2(1, 32) = 3.23$, $p = .08$. The interaction between language context, verb type and motion type was significant in the participant analysis, $F1(1, 32) = 4.08$, $p = .05$, while the interaction between motion type and language context was significant in the item analysis, $F2(1, 32) = 6.71$, $p = .01$.

To further elucidate these findings, two 2 (verb type) X 2 (motion type) ANOVAs were conducted at each level of language context. The first indicated that there was no main effect of motion type ($p > .05$) in the language-absent condition (the effect of verb type had no bearing in this condition, as there was no spoken sentence accompanying the films). In the language-present condition, however, there was a significant main effect of motion type, $F1(1, 14) = 9.13$, $p = .009$, $F2(1, 33) = 21.27$, $p < .0001$, such that SOTs were shorter in the towards condition ($M1 = 1205.4$, $SD1 = 555.8$, $M2 = 1181.7$, $SD2 = 602.7$) than in the neutral condition ($M1 = 1331.9$, $SD1 = 562.3$, $M2 = 1369.9$, $SD2 = 695.6$). Verb type, however, did not produce a significant main effect, $F1(1, 14) < 1$, $p = .75$, $F2(1, 33) < 1$, $p = .39$, nor was the interaction significant ($p > .05$).

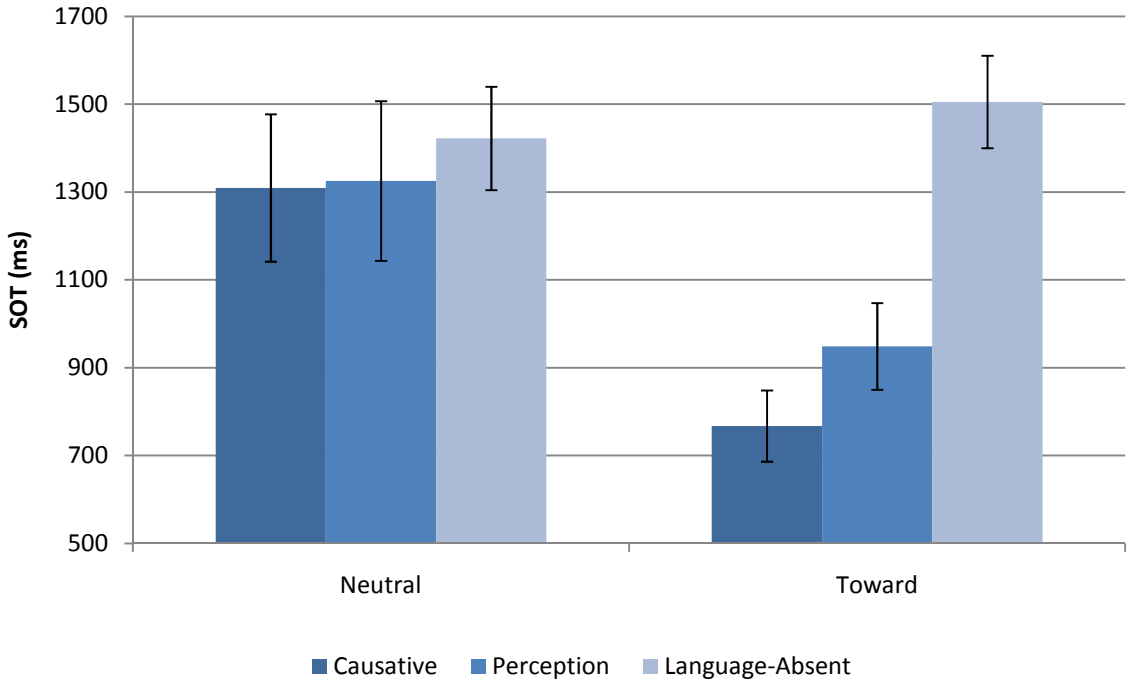


Figure 7. Saccade onset time (SOT) ± SE as a function of language context/verb type and motion type, computed by items. Note that the two language-absent conditions group together the two verb conditions (which represented a false separation in the main ANOVA). The results reported take this into account. The main analysis compared the means of these groups using a 2 (language context) X 2 (verb type) X 2 (motion type) ANOVA. Analyses showed that the difference between the two verb types in the language-present-toward condition was only marginally significant, $t(16) = -1.42$, $p = .09$. In addition, in the toward condition, the absence of language led to slower SOTs.

This main ANOVA thus indicates that in the absence of language, SOTs are on the whole longer than in the presence of the spoken sentence (a significant main effect of language context was obtained but not interpreted due to the significant interactions), as predicted. Thus, the event-related utterance does serve to constrain eye movements to the object being named. However, verb-specific constraints do not moderate this effect, and fail to trigger faster eye movements, which replicates the result found in Experiment 1. In addition, as found in the analysis on early post-disambiguating point saccades, it appears that the information conveyed by the motion context fails to moderate the effect of language context when no spoken sentence is present. Thus, the visual and linguistic contexts do incrementally influence each other in interpreting the unfolding event, as evidence by focal attention.

Six planned comparisons were also conducted to test more specific hypotheses. These consisted of a series of one-tailed paired *t*-tests, conducted both by participants (*t1*) and by items (*t2*). First, we tested whether verb type had an effect in the language-present condition. We hypothesized that the causative and perception conditions would differ significantly in the language-present group, such that the causative condition would have a lower mean SOT than the perception condition. A one-tailed paired *t*-test was conducted to that effect, but did not reveal a significant difference, $t1(14) < 1$, $p = .49$, $t2(16) < 1$, $p = .33$, contrary to our prediction, but consistent with the ANOVA conducted above. Thus, perhaps due to the relatively small sample size in the language-present condition, verb effects failed to emerge, indicating that even without the away condition and with the introduction of distractor trials, the sensitivity to verb-specific information in dynamic scenes is not augmented.

Second, to examine the difference between causative and perception verbs without the moderating effect of apparent motion, we compared the causative-neutral and perception-neutral groups in the language-present condition using a one-tailed paired *t*-test, with the expectation that the causative-neutral condition would have shorter SOTs than the perception-neutral condition. A one-tailed paired *t*-test failed to lend support to this notion, with results showing a non-significant difference, $t1(12) < 1$, $p = .25$, $t2(16) < 1$, $p = .47$. We found the same result in Experiment 1, which strengthens the notion that in the absence of agent motion, verb effects do not occur.

Third, we examined the difference between causative-towards and perception-towards groups in the language-present condition, as it was expected that this motion context might serve to increase the speed at which saccades were initiated due to the semantic consistency between the linguistic and visual contexts. We expected that SOTs would be lower in the causative-towards than in the perception-towards condition. A one-tailed paired *t*-test marginally supported this notion in the item analysis, $t2(16) = -1.42$, $p = .09$, but not in the participant analysis, $t1(14) = -1.00$, $p = .17$. This evidence for the effect of verb class in the presence of motion is weak compared to that found in Experiment 1, although again, the small sample size may have failed to produce a more significant results. Nevertheless, the difference did occur in the predicted direction.

Finally, and most importantly for the purposes of this experiment, we compared the language-absent-toward, language-present-neutral, and language-present-towards groups to disentangle the effects of the linguistic context and the visual context in guiding eye movement behaviours. Figure 8 shows these comparisons in graphical form. The language-present-toward group served as the "experimentally confounding" group, as it

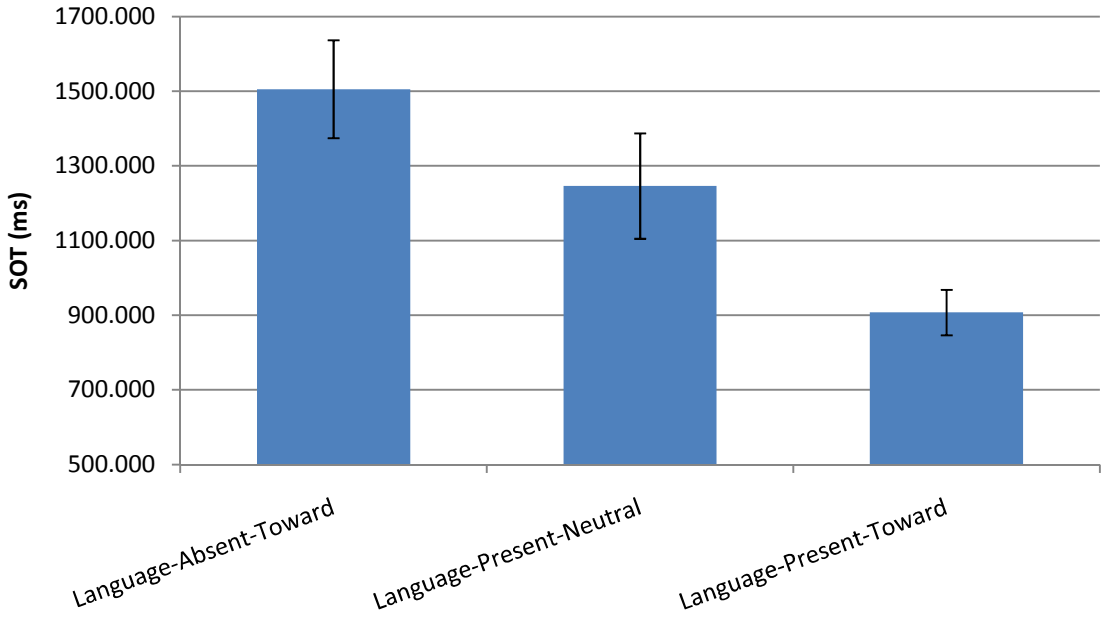


Figure 8. The relative contribution of motion context and language context. The language-absent-toward, language-present-neutral, and language-present-toward conditions were compared to disentangle the effects of language and motion context. Results showed that language context has a stronger relative contribution than the motion context in guiding eye movement behaviour. The combination of both constraining conditions (language and motion) produces the shortest SOTs.

combines the effects of both the linguistic and visual contexts. The other two groups served as controls; one without language but with motion, and the other without motion but with language.

To that end we first compared the two controls, the one in which language was present but motion was absent (the language-present-neutral condition), and the one in which language was absent and motion was present (the language-absent-toward condition), to determine which provided the stronger context. If the language-absent-toward condition were to produce shorter SOTs than the language-present-neutral condition, then the visual context could be said to have a stronger effect than the linguistic context. On the other hand, if the language-present-neutral condition were to produce shorter SOTs than the language-absent-toward condition, then the linguistic context could be said to have a stronger effect than the visual context. We predicted that the linguistic context would have a stronger effect. A one-tailed *t*-test did show a significant difference between the two in the participant analysis, $t1(32) = -1.98$, $p = .03$, and a one-tailed paired *t*-test showed a tendency in the item analysis, $t2(16) = 1.33$, $p = .10$, with the language-present-neutral group ($M1 = 1203.2$, $SD1 = 416.3$, $M2 = 1246.0$, $SD2 = 582.0$) having shorter SOTs than the language-absent-toward group ($M1 = 1497.6$, $SD1 = 440.0$, $M2 = 1505.4$, $SD2 = 539.8$). Therefore, as expected, the linguistic context in the absence of motion led to faster saccade onset times than the visual context in the absence of language.

Next, we compared the two control conditions to the confounding condition. First, to test the hypothesis that the motion context confers an advantage to the constraints provided by the language context, we compared the neutral and toward groups

in the language-present condition, expecting the latter to produce shorter SOTs. A one-tailed paired *t*-test confirmed this hypothesis, $t1(14) = -2.09$, $p = .03$, $t2(16) = -2.07$, $p = .03$, and showed that saccades were launched more quickly in the language-present-toward condition ($M1 = 906.1$, $SD1 = 286.9$, $M2 = 907.3$, $SD2 = 251.1$) than the language-present-neutral condition. Thus, in the presence of language, agent motion toward the target object leads to faster saccades, as previous analyses have shown.

Second, to test the hypothesis that the language context confers an advantage to the constraints provided by the motion context, we compared the toward groups in the language-present and language-absent conditions. If the language-present-toward condition produces shorter SOTs than the language-absent-toward condition, then the spoken language confers an advantage to the motion context. In this case, the visual and linguistic contexts can be said to have an additive effect. On the other hand, if the language-absent-toward condition produces shorter SOTs than the language-present-toward condition, then the spoken language actually introduces a competition for cognitive resources, and hinders the search for the target object in the scene. We expected that the presence of language would in fact lead to shorter SOTs in the toward condition; a one-tailed *t*-test in the participant analysis supported this notion, $t1(32) = -4.49$, $p < .0001$, as did a one-tailed paired *t*-test in the item analysis, $t2(16) = 4.71$, $p = .0001$. The language-present-toward condition did in fact have a lower mean than the language-absent-toward condition, which indicates that the presence of language confers an advantage to the motion context.

This set of planned comparisons was able to directly compare the relative contribution of the language and dynamic visual contexts. While the presence of

language or motion on its own is able to trigger faster eye movements than without, it is the language context that does so more quickly. The results point to the primacy of the linguistic context in guiding eye movement behaviour over the motion context. However, it appears that these two effects are in fact additive, such that in combination they produce the fastest SOTs of all, indicating that each is able to moderate the effects of the other. These findings suggest that the linguistic and visual processing systems do interact to produce an interpretation of the unfolding event, and that the visual-attentional system seeks out the object that is implicated in that event. Given the relatively late initiation of these eye movements, it can be said that they exist to confirm the interpretation being constructed, rather than to anticipate it.

This is the first study with dynamic scenes to date that has directly tested the assumption of the visual world paradigm that situated language processing can in fact be measured by eye movement behaviour. Given that the visual-attentional system is influenced by the presence of language, this basic assumption and the body of research predicated upon it is valid. Despite the fact that verb effects on the whole have failed to emerge within the dynamic scenes our research has used (including Di Nardo, 2005), it is not because the visual complexity (scene realism, dynamic motion) of these scenes entirely dominate the visual-attentional system. Rather, it seems the opposite is true: when directly pitted against each other, it is the linguistic context that triggers the fastest saccades, not the motion context. Thus, saccades are not driven entirely by the perceptual features of the scene (bottom-up processing), but rather the conceptual representations of the unfolding event and the entities encoded therein (top-down processing). In this sense, it can be said that the visual-attentional system, as measured

by visual search patterns, preferentially relies upon top-down cognitive factors in determining fixation location (Henderson & Ferreira, 2004), although this process is further enhanced by bottom-up features that corroborate the interpretation.

Whether the elimination of the away condition and the introduction of the filler movies aided in reducing participants' ability to detect the experimental conditions is less clear. We removed the away trails on the hypothesis that the presence of these semantically inconsistent films may have caused an undue reliance on agent motion (or lack thereof) for interpreting the scene's event meaning. Therefore, saccade onset time should have been lower in the language-present condition in this experiment than in Experiment 1; indeed, the average SOT was approximately 150 ms faster. In addition, compared to the neutral and toward conditions only of Experiment 1, Experiment 2 produced SOTs that were approximately 175 ms faster. Thus, it does seem that participants were somewhat quicker in initiating saccades to the target object in this experiment, and that this may have been a result of a combination of both methodological changes (elimination of away trials, introduction of distractors). Nevertheless, these changes did not allow verb effects to emerge in the language-present condition, and further adjustments to the stimuli may be required to enhance the visual-attentional system's sensitivity to verb-specific constraints.

Overall, the results of Experiment 2 make two important contributions. First, they replicate many of the same findings as Experiment 1. On the whole, the motion context consistently has a significant effect, such that the toward condition usually biases eye movements toward the target object when compared with the neutral condition. In addition, wherever the language context was present, verb effects failed to emerge. Thus,

the dynamic visual world paradigm emphasizes the primacy of the motion context over the *verb* context, but not the *language* context. These results fail to support previous research that indicates verb-specific constraints can guide eye movements (e.g., Altmann & Kamide, 1999; Boland, 2005). Nevertheless, these results do appear to be consistent in our studies of situated language processing in dynamic scenes.

In addition, the second major contribution of this study is that it shows that the presence of language does indeed alter the pattern of eye movements across the dynamic scene, despite the lack of verb effects. First, the absence of language leads participants to fixate the target object more often before the disambiguating point. After that point, these fixations occur less often than when there is language, but only when there is no biasing motion context. In other words, if the agent moves toward the target object, then it receives just as many fixations as when there is spoken language. In addition, early saccades following the disambiguating point clearly show that the presence of language increases the number of saccades occurring by the end of the utterance of the noun, although we did not find any evidence of anticipatory eye movements. Interestingly, the presence of language does not increase the duration of fixations to the target object, even after their utterance, but it does cause those fixations to occur more quickly. Finally, saccades were triggered more quickly in the presence of language, and this linguistic context, when combined with motion, led to the fastest SOTs. Thus, these results provide evidence that not only does the presence of language produce a unique pattern of eye movements, but that the linguistic and visual contexts are incrementally integrated to construct an interpretation of the unfolding event.

Given that the presence of language in the dynamic visual world does in fact make a significant contribution to programming the location and timing of saccadic targets, we asked whether making the linguistic context even more salient would not only have stronger effects on these measures, but would also allow verb effects to emerge. To that end, we designed the materials in Experiment 3 to contrast two versions of the initial patch clause: one that was semantically restrictive, or predictive of, the target object, and one that was not. We expected that by creating patch clauses that made indirect reference to objects or events that semantically "primed" the target object, the visual-attentional system would be strongly biased toward any linguistically constraining cues present in the verb's representation.

Experiment 3

The main purpose of this experiment was to determine whether strengthening the linguistic context through the use of semantically restrictive/predictive initial patch clauses would alter participants' scan paths through the scenes. While the previous experiment demonstrated that the presence of spoken language in the visual world paradigm does contribute to the control of visual fixation, it appears that under the conditions tested in Experiment 1 and Di Nardo (2005), verb-specific constraints help select the domain of reference (saccades to target referent of the grammatical object) only when these match the dynamics of the scene—i.e., agent motion towards the target. While the contrast in dynamic motion of the agent seems to be more salient than the contrast between verb class, it is clear that the overall presence of scene-related spoken language does strongly influence visual search patterns. Thus, by increasing the salience of that linguistic stream early during its utterance, the visual-attentional system should

become more clearly attuned to the linguistic properties of that utterance, particularly those belonging to the verb.

To test this hypothesis, we created two versions of each initial patch clause: one that was semantically non-restrictive, and one that was semantically restrictive. As part of the design of this experiment, a normative study was conducted to collect information used to generate these clauses (described in the Method section below). The restrictive patch clause was designed to bias the linguistic utterance toward a more constrained interpretation of the scene's event meaning, which would therefore constrain the visual domain of reference, particularly post-verb. The non-restrictive clause, on the other hand, tended to make more general reference to the scene's unfolding event, and therefore did not create an expectation of the involvement of the target object. By introducing a more constrained interpretation of the scene's event meaning from the beginning of the linguistic utterance, the incremental interpretation being constructed by linguistic and visual systems should serve to bias visual search mechanisms toward a confirmatory search pattern that occurs more quickly.

Given that the majority of visual world studies have employed visual stimuli that are non-complex and lack dynamic features (the *ersatz* scenes discussed by Henderson & Ferreira, 2004), information conveyed by the verb was relatively more salient and therefore able to affect visual-attentional mechanisms more directly. Introducing a more salient linguistic context is therefore somewhat akin to matching the increased salience of the visual context used in our research. We expected this modification to result in two notable changes: one, that SOTs should be faster in the restrictive condition than the non-restrictive condition, both in the analysis of early post-verb saccades (those that occur by

the offset of the noun) and in the main analyses (of all post-verb saccades to the target object); and two, that more reliable verb effects should emerge across both motion contexts. In addition, we expected participants to fixate the target object for longer durations in the restrictive condition, particularly before the verb onset. Finally, we hypothesized that anticipatory effects would fail to emerge, but that by the offset of the noun, a higher number of saccades toward the target object should have been initiated in the restrictive condition than in the non-restrictive condition. All other experimental variables and conditions were maintained from Experiment 2 (e.g., projection onto the large screen, the elimination of the away trials, and the inclusion of filler trials).

Method

Participants. Forty-seven participants took part in this study, all Concordia University students, and all native speakers of English. None of these participants had taken part in the earlier experiments, nor in the normative studies. There were 40 females and seven males, ranging in age from 17 to 52 years. Data from all 47 participants were retained in the analyses. Inclusion and exclusion criteria were the same as in Experiment 1. All participants received course credit for their participation.

Materials and apparatus. The film clips used in this experiment were identical to those used in Experiment 2. However, significant changes were made to the initial patch clauses of each of the sentences, based on the results of a normative study we carried out to elicit potential event representations (see Appendix E for the list of sentences). The purpose of this normative study was to gather information from participants on their beliefs about what is going on in the scenes employed in Experiments 1-3 (current and likely future events). The information about current and

future events obtained from this study were used provide a range of possible representations that could be considered more or less semantically restrictive to the target object, i.e., made indirect reference to its involvement in the upcoming event described by the main clause. These used to construct the initial patch clauses employed by this experiment. The goal was to collect event representations that were both predictive of the involvement of the target object, and those that were not.

The norms were collected from twenty-five participants drawn from the Concordia University student community, none of whom had taken part in any of the other experiments. There were 20 females and five males, ranging in age from 18 to 38. We presented the movies used in Experiment 1 (one from each of the three motion contexts) without sound, on an iMac G4 17" computer screen using a Microsoft PowerPoint slideshow. These were distributed among three lists and presented in pseudo-random order. Each trial began with the fixation cross for 2 s, followed by the onset of the movie clip.

Participants were instructed to watch each movie clip, starting by fixating on the cross, and to then answer a series of questions about each in a booklet provided (see Appendix G for the instructions given to the participants). They were told that the last frame of each movie would remain on the screen in order to help them in the answering of these questions. Participants were asked to answer the questions as quickly as possible, using the first thoughts that came to mind, and to use complete sentences. They were also warned that some of the questions may sound odd given the context of the film, but to do their best in answering without needing to provide the "right" answer.

The questions participants were asked to answer were: (1) *What do you think is going on in this scene?* (2) *What do you think will happen next?* (3) *Notice the [target object] on/in [location within scene]—what do you think the [agent] will do with it next?* (4) *What is a typical function of [target object]?* (5) *What is typically made with [target object]?* and (6) *What is a typical activity that involves [target object]?*

The first three questions were designed to elicit current and future event descriptions. The third question in particular was included in case the target object was not specifically named in the second question. Questions 4-6 were designed to elicit typical features of the target object, which could assist in the construction of the patch clauses. For example, for the target object *egg*, typical responses to the last three questions included, "to eat," "omelettes," and "cooking."

Once responses were collected, each sentence clause (for the first three questions) was coded as an event structure, with predicate and arguments listed (e.g., [*break* [*the cook, eggs*]]). These predicate/argument combinations were then used in assisting the creation of the patch clauses for the current experiment, with event structures that either were more semantically predictive of the target object or less predictive being preferred (one for each of the two conditions).

The main clause and final patch clause remained identical to maintain experimental constancy after the disambiguating point. In the non-restrictive condition, the initial patch clause made little or no reference to the target object; in this example, *the egg*. However, in the restrictive condition, the initial patch clause was more semantically related to the target object. For example, the two clauses contrasted were *After pouring the flour into the bowl* (non-restrictive) and *In order to make the omelette* (restrictive).

The *omelette* reference was based on the normative study information (see Appendix E). The non-restrictive clause, on the other hand, does not imply which particular event (and its associated object) might occur next.

These sentences were recorded by a female research assistant who produced the spoken sentences for the filler trials in Experiment 2 (also used in the present experiment). Sentences were synchronized with the same movies used in the previous experiments in the same manner described above (Experiment 1). Given the two versions of each patch clause, two verb types, two motion contexts and 17 scenes, a total of 136 unique movie/sentence combinations were produced, with eight versions of each scene distributed across eight lists.

In addition to these 17 movie combinations, the same six filler trials used in Experiment 2 were included. Thus each participant was exposed to a total of twenty-three movies. As in Experiments 1 and 2, the film clips were projected onto a large screen using the same projector, at the same distance and dimensions. All other materials and apparatus were identical.

Procedure. The procedure employed in Experiment 3 was identical to Experiment 1.

Analyses. The same sets of analyses were used as in Experiment 2, including the use of language context as a third independent variable where appropriate—except that in the present experiment semantic context was a within-subjects variable, so that ANOVAs were not mixed- factor. Unless otherwise indicated, all hypotheses are the same as in Experiment 1. Any modifications to the hypotheses for each set of analyses are presented below.

Results and Discussion

As in the previous two experiments, a short cued recall test was given at the end of the experiment to ensure participants paid attention during the trials. Quiz scores ranged from 50% to 100% (M = 96%, SD = 9.74%). The participant who scored at chance was retained in the analyses, given the large proportion of data that was missing overall (reported below).

We first performed a manipulation check to examine the effect of the three independent variables on the proportion of trials (by items) where the participant looked at the target object before the verb onset. We did not expect a difference based on verb type or motion type as these only became apparent after the disambiguating point. However, we did expect a main effect of semantic context; in the condition with more semantically restrictive initial patch clauses, we would expect a "priming" effect in which visual attention would be guided toward the target object early during scene processing.

As expected, neither verb type nor motion type had a significant effect on the proportion of trials in which the participant initiated a fixation to the target object prior to the disambiguating point ($p > .05$). However, as hypothesized, semantic context did have a significant main effect, $F(1, 16) = 3.95$, $p = .004$, such that there was a larger proportion of saccades to the target object prior to the disambiguating point in the restrictive condition (M = .543, SD = .272) than in the non-restrictive condition (M = .412, SD = .322). In other words, the processing of the *omelette* patch clause led to more saccades to *eggs* prior to its utterance. This confirms that the introduction of semantically restrictive patch clauses may have contributed to the building of an event meaning that includes the participation of the target object as one of the role fillers. In fact, these pre-verbal

fixations can be considered anticipatory in the sense that neither the verb nor the noun referent have yet been uttered, and the visual system is attempting to locate likely candidates for the unfolding event based on the current interpretation of its meaning gleaned from scene composition and the initial linguistic utterance. Therefore, the manipulation of semantic context did have its intended effect; the analyses reported below further explore how eye movements were influenced by this manipulation.

Missing data. The proportions of missing data from the same three sources as in Experiments 1 and 2 were computed. The first source of missing data was due to corrupt data, such as a system crash, poor calibration or, more frequently encountered, drift caused by head movements. Out of the 789 trials presented to participants, 272 (34.5%) were lost due to corrupted data. These trials were distributed evenly across the various experimental conditions (no significant main effects).

The second source of missing data was trials in which participants never fixated the target object after verb-onset. Eighty (10.1%) such trials were recorded. A repeated-measures 2 (verb type) X 2 (motion type) X 2 (semantic context) ANOVA ($N = 47$) was conducted (computed by participants) in order to examine whether these factors had an effect on the proportion of trials where participants did not launch a saccade to the target object after verb-onset. There were no significant main or interaction effects between the three independent variables ($p > .05$), suggesting that these trials were evenly distributed across all eight conditions.

Another test was conducted to examine the possible cause of this source of missing data. A Chi-Square test was used to determine whether having looked at the target object before verb onset might have caused participants not to look at the target

object after this point. Results failed to show a significant effect; $\chi = .02$, $p = .90$. Thus, the previous fixation (and therefore encoding) of the target object did not preclude its fixation after the disambiguating point.

Trials in which participants happened to be fixating the target object at verb-onset constituted the third source of missing data, as these trials had to be excluded from any analyses examining the effect of verb type on subsequent eye movement behaviour. This occurred in 81 (10.3%) of the trials. A 2 (semantic context) X 2 (verb type) X 2 (motion type) repeated-measures ANOVA ($N = 47$) was conducted to examine the effect of these factors on the proportion of these trials (by participants). We expected that there would be a higher number of such trials in the restrictive condition than the non-restrictive condition, because the initial patch clauses in the restrictive condition were designed to be predictive of the target event and by implication, the target object. However, we did not expect to find a main effect of either verb type or motion type, as these were not differentiable prior to the disambiguating point.

The results indicated that the interaction between semantic context and motion type was significant, $F(1, 46) = 7.33$, $p = .009$, as was the main effect of verb type, $F(1, 46) = 6.92$, $p = .01$, contrary to our prediction. Further analyses exploring the interaction indicated that the effect of motion type was significant in the restrictive condition, $F(1, 46) = 5.68$, $p = .02$, but not the non-restrictive condition ($p > .05$). Given that neither the motion nor verb contexts were active prior to the disambiguating point, it is unclear why these effects were significant. In addition, the failure of the effect of semantic context to reach significance is contrary to our hypothesis. However, these results do demonstrate

that the pattern of trials in which the participant was already looking at the target object at verb-onset was not evenly distributed among the different conditions.

In summary, the pattern of missing data attributable to two of the three sources (trials in which data was corrupted or participants never looked) was randomly distributed across the eight conditions. In addition, because such a large proportion of data was missing, the analyses reported below used condition means to replace any missing cells. Analyses are reported both by participants (*F1*) and by items (*F2*).

Effect of semantic context on fixations to target objects. In order to examine the effect of semantic context on eye movement behaviour toward the target object, we correlated this variable with the time spent fixating the target object before the verb onset, the time after this point and the total fixation time. We expected that there would be a significant positive correlation, such that in the restrictive condition, the target object would be fixated for longer intervals, particularly prior to the verb onset. A one-tailed point biserial correlation ($N = 449$) was computed, which indicated that there was only a significant correlation between language context and the pre-verb fixation times, $r = .10$, $p = .01$, as predicted. The other two correlations were not significant (after: $r = -.005$, $p = .54$; total: $r = .04$, $p = .19$), contrary to our hypothesis. This suggests that the target object was fixated for longer durations when the initial patch clause was more semantically restrictive than when it was not, even though the object was not directly referred to. This again supports the notion that our manipulation did bias interpretation toward an event involving the target object—that is, semantic context did enhance attention to the target object.

Analysis of early post-verb cumulative saccades to the target object. The effects of sentence point (at verb offset, noun onset and noun onset), semantic context, verb type, and motion type on the cumulative proportion of saccades to the target object after verb-onset was analyzed using a repeated-measures 3 (sentence point) X 2 (semantic context) X 2 (verb type) X 2 (motion type) ANOVA. We expected that there would be main effects of, or a significant interaction between at least two of the four factors, such that the difference between the causative and perception, and toward and neutral conditions would increase as the sentence unfolded, as hypothesized in the prior two experiments. In addition, we also expected the difference between the non-restrictive and restrictive conditions to increase between the offset of the verb and the offset of the noun, as the initial segment of the utterance should have biased the visual-attentional system toward the target object even at this early point.

The results (which are plotted in Figures 9 and 10; see also Tables 9 and 10 for the ANOVA tables relevant to this analysis) confirmed our hypotheses, such that the four-way interaction was significant in the participant analysis, $F1(2, 84) = 9.24$, $p = .0002$, and marginally significant in the item analysis, $F2(2, 32) = 3.01$, $p = .06$. The main effect of sentence point was also significant, $F1(2, 84) = 93.00$, $p < .0001$, $F2(2, 32) = 100.39$, $p < .0001$, as were the main effects of language context, $F1(1, 42) = 4.50$, $p = .04$, and motion type, $F1(1, 42) = 4.20$, $p = .05$, but only in the analysis by participants. The main effect of verb type was not significant ($p > .05$).

To explore this interaction, tests of the simple effects of semantic context, verb type, and motion type across all levels of sentence point were conducted. The first of these tests indicated that there was a marginally significant main effect of semantic

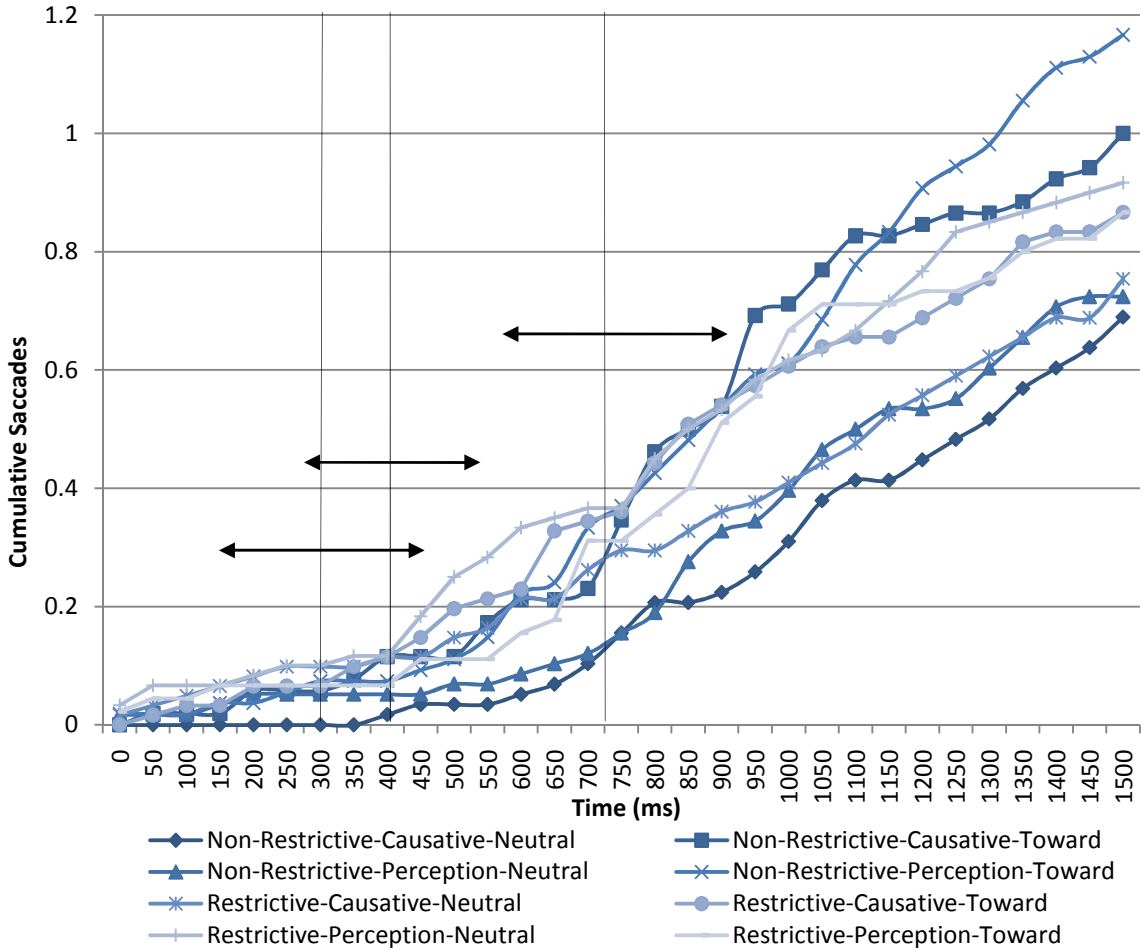


Figure 9. A plot of the mean cumulative average number of fixations (by participant) to the target object after verb-onset. Each line refers to a single condition, and each point to one 50-ms bin. The origin of the X-axis refers to the verb-onset, and the three vertical lines mark the temporal boundaries of the verb and noun (average onset and offset). The double-headed horizontal arrows on each boundary indicate the range of onsets and offsets at the points in time relative to the verb-onset, taking into account the variable lengths of each of the sentence segments (i.e. verbs and noun phrases). The point at which each of the coloured lines (referring to cumulative fixations for each condition) intersects with the three critical sentence points (verb-offset, noun-onset and noun-offset) were computed and compared in an ANOVA.

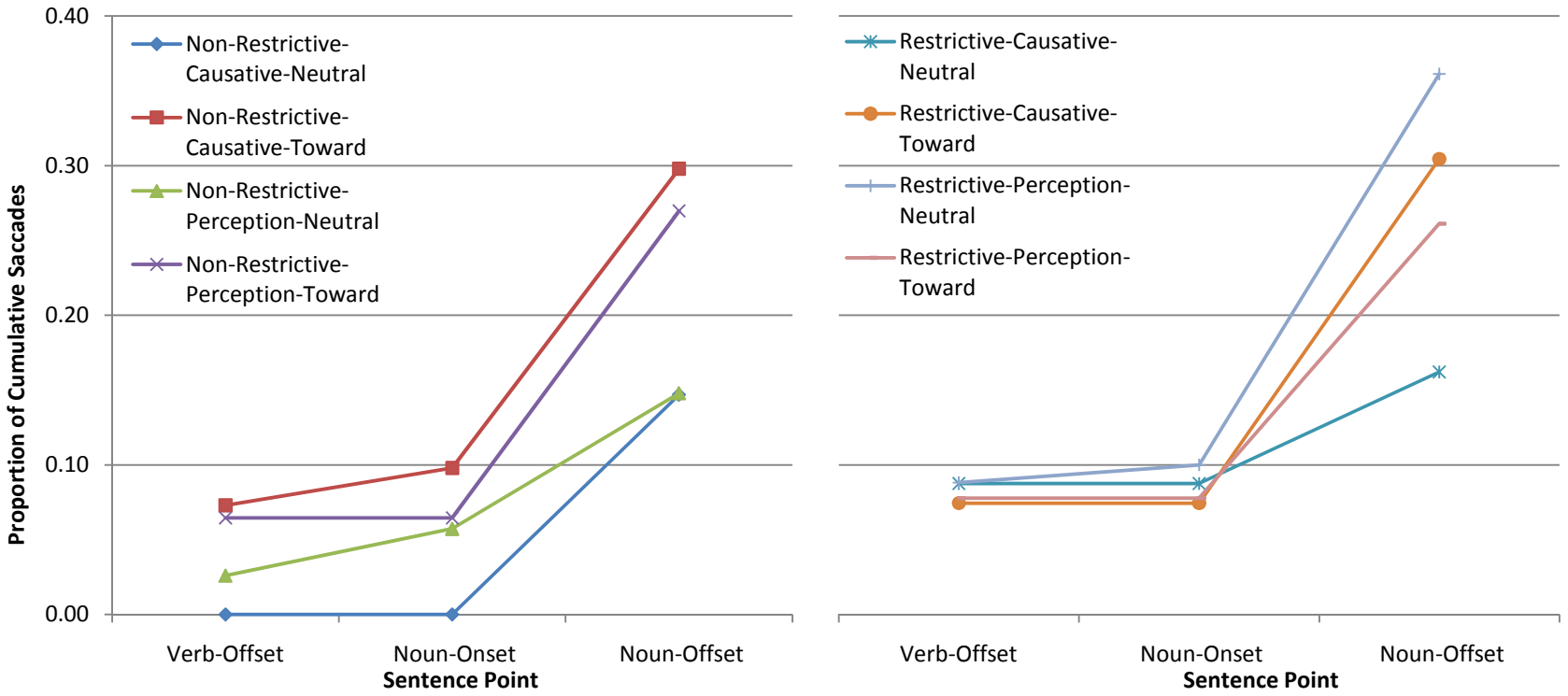


Figure 10. Mean number of cumulative saccades towards target object at each of the three critical sentence points, verb-offset, noun-onset and noun-offset, separated by language context (non-restrictive on the left, restrictive on the right). A cursory examination of the graph shows that the four towards conditions are among the highest means at noun-offset, exhibiting very early post-verb effects. However, the highest mean is for the restrictive-perception-neutral group; in the absence of semantically restrictive information in the verb, and of motion, the initial patch clause appears to have its greatest effect post-verb.

context at the verb offset in the participant analysis, $F1(1, 42) = 5.90$, $p = .02$, and a tendency toward significance in the item analysis, $F2(1, 16) = 3.33$, $p = .09$, such that the restrictive group (*M1* = .141, *SD1* = .281, *M2* = .146, *SD2* = .207) had a higher mean proportion of cumulative saccades toward the target object than the non-restrictive group (*M1* = .097, *SD1* = .235, *M2* = .104, *SD2* = .193).

At noun-onset, however, the effect of semantic context failed to reach significance, $p > .05$. Instead, the interaction between verb type and motion type was marginally significant in the participant analysis, $F1(1, 42) = 3.60$, $p = .06$, while there were no significant main effects or interactions in the item analysis (all $p > .05$). A test of the simple effects of verb type at each level of motion context revealed that verb type was not significant in the toward condition, $F1(1, 42) < 1$, $p = .38$, but was significant in the neutral condition, $F1(1, 42) = 4.52$, $p = .04$, such that the perception condition (*M1* = .068, *SD1* = .172, *M2* = .079, *SD2* = .124) had a higher mean proportion of cumulative saccades than the causative condition (*M1* = .030, *SD1* = .104, *M2* = .044, *SD2* = .105). This indicates that at the noun-onset, in the absence of agent motion, more saccades were launched following the less restrictive perception verb than the causative verb, which is contrary to our hypothesis.

Finally, at the offset of the noun, the interaction between semantic context, verb type and motion type was significant in the participant analysis, $F1(1, 42) = 7.00$, $p = .01$, while there were no significant main effects or interactions in the item analysis. To further explore this interaction, tests of the simple effects of semantic context and verb type at each level of motion type were conducted. These indicated that in the neutral condition, the interaction between semantic context and verb type was significant, $F1(1,$

42) = 10.40, p = .002. Further tests showed that verb type only had a significant main effect in the restrictive-neutral condition, $F1(1, 42) = 9.42$, p = .003, with the perception condition ($M1$ = .382, $SD1$ = .376) having a greater number of cumulative saccades than the causative condition ($M1$ = .162, $SD1$ = .280). In the toward condition, neither verb type nor semantic context had a significant main effect, nor was the interaction significant ($p > .05$). These results therefore indicate that at the end of the noun's utterance, the only significant difference found was between the restrictive-perception-neutral and restrictive-causative-neutral conditions. Following the semantically restrictive initial clause, and in the absence of motion, it appears that perception verbs lead to more saccades than causative verbs. This is contrary to what might be expected, but in keeping with the results found at the noun-onset.

To further test our hypotheses, a series of four planned comparisons were carried out (both by participants, $t1$ and by items, $t2$). First, we hypothesized that the toward group would exhibit significantly more cumulative saccades than the neutral group across all sentence points. This hypothesis was not supported: a one-tailed paired t-test failed to reach significance, $t1(42) < 1$, p = .35, $t(16) = -1.00$, p = .16. This is in contrast to the findings of Experiment 1, but consistent with Experiment 2. There are two possible explanations for this finding: one, the elimination of the away condition might have reduced sensitivity to the motion context; or two, the variation in linguistic context (both its absence in Experiment 2 and the variation in initial clauses in Experiment 3) might also have rendered the visual-attentional system less sensitive to the motion context, at least at this early stage of post-verb processing.

Second, we tested the hypothesis that the restrictive group would have significantly more cumulative saccades than the non-restrictive group across all sentence points using a one-tailed *t*-test. This hypothesis was marginally confirmed in the item analysis, $t2(16) = -1.61$, $p = .06$, but not in the participant analysis, $t1(42) = -1.22$, $p = .12$. Thus, between the offset of the verb and the offset of the noun, there was a tendency for the more semantically restrictive initial clause to yield more saccades to the target object.

Third, the hypothesis that the causative-neutral condition would have a higher proportion of cumulative saccades than the perception-neutral condition across all sentence points and language contexts was marginally supported. A one-tailed paired *t*-test reached marginal significance in the item analysis, $t2(16) = -1.52$, $p = .07$, but did not reach significance in the participant analysis, $t1(40) < 1$, $p = .17$. This indicates that in the absence of motion, verb effects do not reliably emerge during the utterance of the main clause. This supports the findings of Experiments 1 and 2.

Fourth, the hypothesis that the causative-towards condition would have a higher proportion of cumulative saccades than the perception-towards condition across all sentence points and language contexts was not supported, $t1(40) < 1$, $p = .79$, $t2(16) < 1$, $p = .71$. Thus, even in the presence of motion, which has previously allowed verb effects to emerge, verb-specific information fails to influence eye movement behaviour.

On the whole, the results of these analyses indicate a pattern that differs somewhat from the results obtained in previous experiments. Most notably, we failed to find verb effects in the toward condition, unlike in the two previous experiments. Because the restrictive semantic context only produced a significant difference at the verb

offset, rather than over the course of the main clause, it seems this variable has effects that are not predictable. However, this measure of eye movement behaviour, the number of cumulative saccades between verb-onset and noun-offset, is a very early and relatively sparse measurement of verb effects. Whether this trend continues with the other measures of eye movement behaviour, in particular the timing of these first saccades, which are more reflexive of verb-semantic interpretation, is examined in the analyses that follow.

Anticipatory eye movements. In order to determine whether anticipatory eye movements occurred, saccade onset times (SOTs) were compared to two other time points in the sentences: the noun-onset and the noun-offset. In addition, we calculated the proportion of trials in which a saccade was launched towards the target object before the onset of the noun. On average, eye movements were initiated 791 ms after the noun-onset in the non-restrictive condition, and 633 ms after the noun-onset in the restrictive condition. In addition, saccades were launched, on average, 525 ms after the offset of the noun in the causative condition, and 686 ms after the offset of the noun in the perception condition. Finally, saccades were launched toward the target object before the onset of the noun in 6.1% of the non-restrictive trials, 9.9% of the restrictive trial, 6.5% of the causative trials, and 9.6% of the perception trials. However, saccades were launched toward the target object before the offset of the noun in 24.4% of the non-restrictive trials, 28.2% of the restrictive trials, 26.6% of the causative trials, and 26.0% of the perception trials. The differences between these groups, at both noun-onset and noun-offset, were not significantly different ($p > .05$), as shown by two 2 (semantic context) X 2 (verb type) repeated-measures ANOVA conducted at the two boundaries of the noun.

In addition, the difference in time between the onset of the noun and SOT was computed and subjected to a 2 (semantic context) X 2 (verb type) X 2 (motion type) repeated-measures ANOVA, in order to determine whether these first saccades were affected by these three factors. Results indicated that the three-way interaction was significant in the participant analysis, $F1(1, 43) = 4.44$, $p = .05$, while the interactions between semantic context and motion type, $F2(1, 16) = 8.79$, $p = .009$, and verb type and motion type, $F2(1, 16) = 7.08$, $p = .02$, were both significant in the item analysis. The main effect of motion type was also significant, $F1(1, 43) = 9.82$, $p = .003$, $F1(1, 16) = 7.12$, $p = .02$, as was the main effect of language context, but only in the participant analysis, $F1(1, 43) = 10.05$, $p = .003$, $F2(1, 16) = 1.76$, $p = .20$. The main effect of verb type was not significant ($p > .05$).

Further analyses at each level of semantic context showed that motion type had a significant main effect in the non-restrictive condition, $F1(1, 43) = 21.62$, $p < .0001$, $F2(1, 16) = 15.96$, $p = .001$, such that the toward condition ($M1 = 646.0$, $SD1 = 423.4$, $M2 = 572.5$, $SD2 = 352.6$) led to faster SOTs following the noun-onset than the neutral condition ($M1 = 811.0$, $SD1 = 567.2$, $M2 = 806.1$, $SD2 = 521.4$). In the restrictive condition, the interaction between verb type and motion type was significant, $F1(1, 43) = 10.33$, $p = .002$, $F2(1, 16) = 6.70$, $p = .02$. Further tests showed that in the restrictive semantic context, verb type only had a significant main effect in the neutral condition in the item analysis, $F2(1, 16) = 3.91$, $p = .05$, $F1(1, 16) = 1.53$, $p = .22$. However, the difference was in the direction opposite to that predicted: here, perception verbs led to faster saccades than causative verbs. The main effect of verb type in the toward condition in the restrictive semantic context was not significant, $p > .05$.

The results of this analysis show that relative to the offset of the noun, saccades are launched more quickly in the toward condition than the neutral condition, when the initial semantic context is non-restrictive. However, when the initial semantic context is restrictive, the effects of motion type are less straightforward given their interaction with verb type: here, verb effects emerge only in the neutral condition. Taken together, these results suggest that the semantic context moderates the effects of verb and motion context in different ways. When it is less restrictive, results similar to the previous two experiments emerge, with only the motion context having an influence in guiding eye movements. However, when it is restrictive and the motion context is neutral, it seems that perception verbs produce an advantage over causative verbs. Thus, the salience of the initial clause can bias eye movement behaviour but only if the motion and verb contexts are non-restrictive.

The next analysis examined the difference in time between the offset of the noun and SOT. The same pattern of results was found. Namely, the three-way interaction in the participant analysis was significant, $F1(1, 43) = 6.28$, $p = .03$, and the interactions between language context and motion type, $F2(1, 16) = 6.18$, $p = .02$, and verb type and motion type, $F2(1, 32) = 8.05$, $p = .01$, were significant in the item analysis. In addition, the main effect of motion type was also significant, $F1(1, 43) = 11.55$, $p = .002$, $F1(1, 16) = 7.96$, $p = .01$, while the main effect of language context was only marginally significant in the participant analysis, $F1(1, 43) = 3.63$, $p = .06$, $F1(1, 16) < 1$, $p = .47$. The main effect of verb type was not significant ($p > .05$).

Further analyses at each level of language context showed that, as before, motion type had a significant main effect in the non-restrictive condition, $F1(1, 43) = 25.78$, $p <$

.0001, $F2(1, 16) = 14.97$, $p = .001$, with saccades being launched more quickly relative to the offset of the noun in the toward condition ($M1 = 277.2$, $SD1 = 416.1$, $M2 = 221.4$, $SD2 = 336.5$) than the neutral condition ($M1 = 453.8$, $SD1 = 535.8$, $M2 = 470.8$, $SD2 = 502.4$). In addition, the interaction between verb type and motion type was again significant in the restrictive condition, $F1(1, 43) = 12.43$, $p = .001$, $F2(1, 16) = 7.68$, $p = .01$. Further tests showed that there was a significant difference between the causative and perception conditions in the neutral motion context, $F1(1, 43) = 7.35$, $p = .008$, $F2(1, 16) = 5.52$, $p = .02$, as well as in the toward condition, but only in the participant analysis, $F1(1, 43) = 4.93$, $p = .03$, $F2(1, 16) = 1.22$, $p = .28$. As at noun-onset, perception verbs had faster SOTs than causative verbs in the neutral condition, whereas in the toward condition, the opposite was true.

Thus, at noun-offset, a pattern similar to that at noun-onset emerged. In other words, in the absence of a restrictive semantic context, only motion effects occur. In the presence of a restrictive semantic context, a more complex pattern is observed: as in the previous analysis, without motion, the perception condition shows an advantage, but with motion, the causative condition leads to faster saccades. Thus, when the initial clause of the spoken sentence is semantically constraining, it seems that saccades are only launched more quickly when both the motion and verb contexts are equally constraining or non-constraining—that is, in the absence of motion, perception verbs lead to faster SOTs than causative verbs, but when motion is present, the opposite is true.

In order to determine whether participants were able to anticipate the target object *before* the agent in the scene reached it in the toward condition, the difference between SOT and the time at which the agent touched the object was computed. It was found that

participants launched a saccade towards the target object 2067 ms before the agent made contact with the object, and as previously found, this indicates that participants launched saccades well before the agent reached the target object, using some combination of visual and linguistic contextual factors. A 2 (semantic context) X 2 (verb context) repeated-measures ANOVA did not reveal a significant main effect of language context or verb type ($p > .05$), suggesting that these factors did not influence the speed at which saccades were initiated toward the target object in the toward motion context.

Target object saliency. The first analysis examined the correlation between target object saliency ratings, pre-verb fixation durations, post-verb fixation durations and total fixation durations. We hypothesized that the saliency ratings would correlate positively and significantly with the amount of time spent looking at the target objects, as more salient objects within the scene should attract more, or longer, fixations. A one-tailed Pearson's correlation ($N = 449$) indicated that target object saliency ratings did not correlate significantly with the time spent looking at the target object before verb-onset ($r = -.003$, $p = .52$), nor the time spent looking after verb-onset ($r = -.04$, $p = .82$), or the total amount of time ($r = -.04$, $p = .80$). This failed to confirm our hypothesis, but is consistent with the findings of Experiments 1 and 2. In addition, we computed these correlations again, separating the restrictive and non-restrictive conditions. We found that in the non-restrictive condition, none of these correlations were significant (total fixation time: $r = -.06$, $p = .20$; fixation time before verb-onset: $r = .03$, $p = .32$; fixation time after verb-onset: $r = -.07$, $p = .13$). The same was true in the restrictive condition (total fixation time: $r = -.02$, $p = .36$; before verb-onset: $r = -.04$, $p = .28$; after verb-onset:

r = -.006, p = .46). Thus, more salient objects failed to receive longer fixations, even prior to the verb-onset when the initial semantic context was restrictive.

Second, the relationship between target object saliency and SOT was computed. We hypothesized that there would be a significant negative correlation, which a one-tailed Pearson's correlation (N = 361) did not confirm (r = - .04, p = .21), which replicates the findings of Experiments 1 and 2. Again, we correlated target object saliency with SOT separately for each semantic context. In the non-restrictive condition, this correlation was not significant (r = -.01, p = .42), nor was it significant in the restrictive condition (r = -.08, p = .16). Because these correlations were not significant, target object saliency was not included as a covariate in the main analysis described below.

Third, the relationship between target object saliency ratings and whether or not the target object was being fixated at verb-onset was examined. We expected that there would be a significant positive correlation, such that the higher the saliency rating, the more likely the target object would be fixated at verb-onset. A one-tailed point biserial correlation (N = 517) was computed, which indicated that there was no significant correlation (r = - .04, p = .82), contrary to our hypothesis, but in line with the previous two experiments. In summary, target object saliency does not correlate significantly with these measures of fixation duration, saccade onset time, or probability of fixation at verb-onset.

Target event saliency. The same set of analyses described above was conducted with target event saliency instead of target object saliency. First, the relationship between target event saliency ratings and the three fixation durations was examined and was

expected to yield significant positive correlations. This hypothesis was only partially supported: a one-tailed Pearson's correlation ($N = 449$) indicated that the amount of time spent looking at the target object before verb-onset did positively correlate significantly with target event saliency ($r = .09$, $p = .03$). However, neither the time spent fixating after ($r = -.08$, $p = .94$) nor the total time spent fixating ($r = -.03$, $p = .55$) were significantly correlated to target event saliency at all. Insofar as the restrictive initial clause contributed to the building of an event representation consistent with the target event, it seems that this likely explains the positive correlation found in pre-verb fixation durations.

Next, the relationship between target event saliency and SOT was computed. We hypothesized that there would be a significant negative correlation, which a one-tailed Pearson's correlation ($N = 361$) did not confirm ($r = .05$, $p = .84$), as found in Experiments 1 and 2. Because of this, target event saliency was not included as a covariate in the main analysis described below.

Finally, the relationship between target event saliency ratings and whether or not the target object was being fixated at verb-onset was examined. We expected that the more predictive a scene was in terms of the target event, the more likely the target object (implicated in the target event) would be fixated at verb-onset. A one-tailed point biserial correlation ($N = 517$) was computed, which indicated that there was a significant correlation ($r = .08$, $p = .04$), which confirmed the hypothesis. Again, this indicates that the restrictive initial patch clause does bias the visual-attentional mechanism toward the target object, such that its probability of being fixated is increased shortly after the utterance of that initial clause (namely, at verb onset). This confirms the notion that

again, the manipulation of the semantic context does influence eye movement behaviour toward the target object.

Main analyses: Saccade onset time. The first analysis examined the effect of verb type and motion type on post-verbal eye movement behaviour, namely saccade onset time (both by participants [*F1*] and by items [*F2*]; due to the large amount of missing data, empty cells were replaced with condition means). We hypothesized that all three independent variables (verb type, motion type and semantic context) would have a main effect on SOT. This hypothesis was partially confirmed, as shown in Figure 11 (see Tables 11 and 12 for the ANOVA tables relevant to the analyses for this experiment). The results of a 2 (semantic context) X 2 (verb type) X 2 (motion type) repeated-measures ANOVA indicated that motion type had a significant main effect, $F1(1, 42) = 11.50$, $p = .002$, $F1(1, 16) = 6.82$, $p = .02$, as did language context, but only in the analysis by participants, $F1(1, 42) = 4.81$, $p = .03$, $F1(2, 84) = 1.65$, $p = .22$. The three-way interaction was also significant in the participant analysis, $F1(1, 42) = 5.64$, $p = .02$, while in the item analysis, the interaction between motion type and language context was significant, $F2(1, 16) = 7.89$, $p = .01$, as was the interaction between motion type and verb type, $F2(1, 32) = 6.50$, $p = .02$. The main effect of verb type was not significant ($p > .05$).

To explore the three-way interaction, two 2 (verb type) X 2 (motion type) ANOVAs were conducted at each level of semantic context. The first indicated that there was a main effect of motion type in the non-restrictive group, $F1(1, 42) = 26.81$, $p < .0001$, $F2(1, 16) = 15.04$, $p = .001$, such that SOTs were shorter in the toward condition ($M1 = 1016.8$, $SD1 = 408.9$, $M2 = 960.0$, $SD2 = 337.6$) than in the neutral condition (*M1*

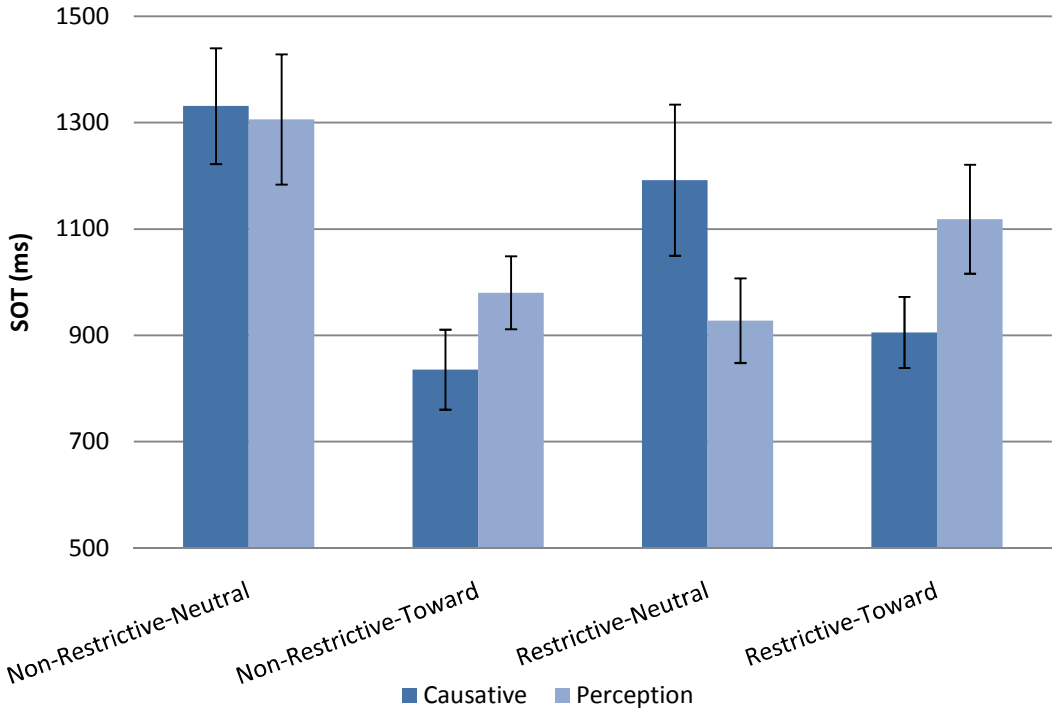


Figure 11. Saccade onset time (SOT) ± SE as a function of language context, verb type and motion type, computed by items. The main analysis compared the means of these eight groups using a 2 X 2 X 2 ANOVA. Analyses showed that the difference between the two verb types in each of the two toward groups was significant in the expected direction. However, the difference between the two verb types in the restrictive-neutral condition was marginally significant in the direction opposite to that hypothesized.

= 1192.1, *SD1* = 541.7, *M2* = 1189.4, *SD2* = 492.9). In the restrictive group, however, there was a significant interaction between motion type and verb type, $F1(1, 42) = 11.61$, $p = .001$, $F2(1, 16) = 6.70$, $p = .02$. Further tests showed that a significant verb effect was only found in the toward condition in the participant analysis, $F1(1, 42) = 9.87$, $p = .002$, with a similar trend in the item analysis, $F2(1, 16) = 3.03$, $p = .09$.

The results of these analyses show that the pattern of results differs between the two semantic contexts. In the non-restrictive context, only motion type affects the speed at which saccades are initiated, whereas verb type does not. However, when the initial semantic context is restrictive, the main effect of verb type emerges, but only in the toward condition.

In order to further explore the three-way interaction, two 2 (semantic context) X 2 (verb type) ANOVAs were conducted at each level of motion type. The first indicated that there was a main effect of semantic context in the neutral group, $F1(1, 43) = 14.52$, $p = .0004$, $F2(1, 16) = 8.08$, $p = .01$, such that SOTs were faster in the restrictive condition (*M1* = 1049.9, *SD1* = 423.4, *M2* = 1060.0, *SD2* = 487.1) than in the non-restrictive condition (*M1* = 1334.3, *SD1* = 608.3, *M2* = 1318.8, *SD2* = 470.6). In the toward group, however, there was a significant main effect of verb type, $F1(1, 43) = 7.89$, $p = .007$, $F2(1, 16) = 6.80$, $p = .02$, such that SOTs were faster in the causative condition (*M1* = 927.0, *SD1* = 365.8, *M2* = 870.6, *SD2* = 290.8) than in the perception condition (*M1* = 1106.5, *SD1* = 431.6, *M2* = 1049.4, *SD2* = 361.0).

Again, the pattern of results differs between the two motion conditions. In the absence of agent motion, saccades are initiated more quickly when the initial semantic context is restrictive, regardless of the semantic constraints of the verb. However, when

the agent does initiate movement toward the target object, semantic context fails to influence eye movement behaviour, while verb class does. In other words, there is no main effect of verb overall, but in the presence of motion, the visual-attentional system is rendered more sensitive to the constraints imposed by the verb.

To more specifically test the hypotheses of this analysis, four planned comparisons were conducted. The first hypothesis was that the causative and perception conditions would differ significantly across all levels of semantic context and motion type, such that the causative condition would have a faster mean SOT than the perception condition. A one-tailed paired *t*-test was conducted to that effect, but did not reveal a significant difference, $t1(42) = -1.16$, $p = .13$, $t2(16) = 1$, $p = .19$, contrary to our prediction.

Second, to compare the two verb types without the confounding effects of the semantic and motion contexts, the causative-neutral and perception-neutral groups were compared. We hypothesized that the causative-neutral condition would have a faster mean SOT than the perception-neutral condition. In the absence of any apparent motion in the scenes, we expected that the causative condition would lead to faster SOTs than in the perception condition. A one-tailed paired *t*-test failed to lend support to this hypothesis, $t1(40) = .21$, $p = .58$, $t2(16) = .46$, $p = .67$. This is consistent with the results of the planned comparisons of Experiments 1 and 2.

Third, to compare the two verb types in the toward condition across both levels of the semantic context, we compared the causative-toward and perception-toward conditions. We expected that the visual context (with the agent moving towards the target object) would aid in the semantic interpretation of the verb, such that SOTs would be

faster in the causative-towards than in the perception-towards condition. A one-tailed paired *t*-test supported this prediction, $t1(39) = -1.90$, $p = .03$, $t2(16) = -2.34$, $p = .01$. Therefore, across both semantic contexts, when the motion context was biased toward the target object, saccades were launched more quickly in the causative than the perception condition.

Fourth, to compare the effects of semantic context without the moderating effects of motion context, we compared the non-restrictive-neutral and restrictive-neutral conditions. In the absence of any apparent motion in the scenes, we expected that the restrictive condition would lead to lower SOTs than the non-restrictive condition. A one-tailed paired *t*-test confirmed this hypothesis, $t1(40) = 2.80$, $p = .004$, $t2(16) = 2.46$, $p = .01$. Therefore, following the utterance of an initial clause that is predictive of the target object, the visual referent of that object receives fixations more quickly than when the clause is less predictive.

Taken together, the results of these planned comparisons generally confirm the hypotheses set out for this experiment. Most notably, the main contribution of the present experiment is the finding that biasing the initial segment of the linguistic stream toward the unfolding event/object representation does trigger faster eye movements toward the target object. In addition, as found in the two previous experiments, when the agent moves toward the target object, the constraints imposed by the verb also direct visual attention toward the target object. This verb effect fails to emerge in the absence of motion, as well as in the overall set of data. In contrast, the findings of the main ANOVA produced a more complex picture. Here, we also found verb effects in the toward motion condition, but only when the initial semantic context is restrictive. When

it is not, only motion type influences the speed at which saccades are initiated. Thus, the initial linguistic stream has a moderating effect on both verb and motion type, the two variables that only become apparent after the utterance of this stream.

The main purpose of this experiment was to determine if creating a more semantically constraining initial clause of the spoken sentence would strengthen the linguistic context relative to the motion context. We expected this to affect eye movement behaviour in two ways: one, by triggering more, or longer, fixations to the target object even prior to the disambiguating point, and two, by allowing verb effects to emerge more consistently. The first general hypothesis was confirmed via several measures. First, the target object received both more and longer fixations prior to the verb-onset in the restrictive than the non-restrictive condition. In addition, at the offset of the verb (i.e., before the noun referent was even uttered), there was a greater proportion of cumulative saccades to the target object in the restrictive than the non-restrictive condition. This difference was not found at the noun's onset and offset, although a planned comparison did reveal a significant difference across all three sentence points. Finally, the main analysis showed that overall, the restrictive semantic context led to faster SOTs overall than did the non-restrictive context.

With regards to the second main hypothesis, that introducing a more semantically restrictive context would allow verb effects to emerge in the various analyses, results were inconsistent. In the main analysis of saccade onset times, verb type failed to have a significant main effect, but rather produced results similar to both the previous two experiments reported here and in Di Nardo (2005). Namely, verb effects were only significant in the toward condition. Furthermore, this was only true in the restrictive

condition; when the initial linguistic utterance was not predictive of the target object, and the agent moved toward the target object, verb type did not lead to a significant difference. Similarly, in the presence of motion across both semantic contexts, verb type also produces a significant main effect.

However, when examining the number of early post-verb saccades, a less consistent pattern resulted. The only verb effects that were found were in the opposite direction to that predicted. More specifically, we found that at the onset of the noun, in the neutral condition, perception verbs led to more saccades than causative verbs. In addition, at the offset of the noun, in the restrictive-neutral condition, perception verbs triggered a greater proportion of saccades to the target object than did causative verbs. While not robust, these findings do represent a departure from the general pattern found thus far, in which verb effects only tend to emerge when the agent moves toward the target object.

While examining the time course of these first saccades relative to the two phonetic boundaries of the noun, an equally inconsistent pattern emerged. Compared to both noun onset and offset, saccades to the target object were launched more quickly in the perception condition than the causative condition, but only in the restrictive-neutral conditions. However, in the restrictive-toward condition, saccades were launched sooner after the noun-offset for causative verbs than perception verbs, which is consistent with our hypotheses. Thus, in this particular set of analyses, it appears that when the initial linguistic stream is semantically restrictive, eye movements are only launched more quickly when both the motion and verb contexts are non-restrictive.

The results of this study make an important contribution to our understanding of situation language processing in dynamic scenes. They indicate that manipulating the salience of the linguistic context by making the initial clause more predictive of the target object can in fact alter the pattern of eye movements. This corroborates the findings of Experiment 2, which tested the ability of the spoken language to guide visual attention toward the target object. Despite the lack of verb effects, which fails to support the findings of other studies (Altmann & Kamide, 1999, in particular), the results of these two experiments clearly show that the overall nature and presence of the linguistic utterance influences visual search patterns. This is true even given the complexity of the scenes employed, which were both realistic and contained dynamic human motion. Thus, in the context of the dynamic visual world paradigm, the failure of verb-specific information to constrain visual reference is not a function of insensitivity to the linguistic stream, but rather to the less salient representations encoded by verb structure.

Furthermore, this study showed that the manipulation of the initial semantic context does not produce anticipatory eye movements. One might have expected the highly constraining combination of a restrictive semantic context, causative verbs and motion toward the target object to led to anticipatory eye movements, but this was not found to be true. However, despite very early linguistic information contained in the initial clause that might serve to create a complete event representation, particularly when integrated with the activation of the set of objects contained within the scene, visual search mechanisms did not seek out the target object soon after the utterance of the verb. Instead, saccades to the target object were launched only after the noun was uttered, again

providing evidence (consistent with our previous study; Di Nardo, 2005) that visual search patterns are confirmatory.

General Discussion

The purpose of the research presented here was to examine how event meanings are built from the contribution of dynamic scene information and spoken language comprehension. More specifically, the study manipulated several visual and linguistic variables in the dynamic visual world paradigm we first introduced (Di Nardo, 2005), in order to determine the relative contribution of these variables to the construction of event meaning. The first experiment sought to test the hypothesis that the lack of verb effects found in Di Nardo (2005) was due to shifts of attention without concomitant eye movements. By projecting the films onto a large screen instead of a computer screen, effectively increasing the visual angle by 72%, any verb-driven shifts to the target object would be more likely to require corresponding saccades. However, we failed to find support for this hypothesis. As found in Di Nardo (2005), verb effects did emerge in the condition where the scene's agent moved toward the target object, but did not emerge more consistently across all motion contexts. Furthermore, no evidence for anticipatory eye movements was found, as saccades were initiated approximately 530 ms (524 ms in the perception condition, 534 ms in the causative condition) after the offset of the noun. Thus, the dynamic, realistic scenes used in both studies failed to replicate the verb-driven, anticipatory eye movements found in studies using *ersatz* scenes, such as Altmann and Kamide (1999) and Knoeferle and Crocker (2007). Instead, the salience of the motion context appeared to more fully control visual fixation patterns.

The purpose of the second experiment was to determine whether, given the lack of verb effects found in Experiment 1 and Di Nardo (2005), the linguistic stream has any ability to drive eye movements toward the target object independently from the visual features of the scene. More specifically, the aim was to explore the relative contributions of visual context and linguistic processes in the control of visual attention as measured by eye movement behaviour. This was accomplished through the manipulation of three variables: the presence *vs.* absence of spoken language, in addition to the previously contrasted verb types and motion contexts (but only the neutral and toward conditions). In addition, we directly contrasted the relative contribution of both the linguistic and visual contexts, as well as their ability to moderate each other's effects. To explore these questions, we compared eye movement patterns across the dynamic scenes with and without the accompanying spoken sentences. We obtained consistent findings that the presence of the utterance does in fact drive visual attention toward the target object, despite the lack of sensitivity to the constraints imposed by the verb. Thus, the lack of verb effects previously found in our studies using dynamic scenes is not due to the visual-attentional system having been entirely dominated by the features of the scene. However, with dynamic scenes, it is likely that the visual-attentional system is only able to attend to the gross aspects of the linguistic stream, such as overall gist or object names, rather than specific syntactic or semantic features such as verb class.

Interestingly, however, this experiment also provided clear evidence that the control of visual fixation preferentially relies on the linguistic stream over the visual context. When directly contrasted, results indicated that verb-specific constraints (in the absence of motion) led to faster saccade onset times than the constraints imposed by the

agent's direction of motion (in the absence of language). Thus, although dynamic scenes tend to dampen the emergence of verb effects when the motion context is not constraining, it is nevertheless sensitive to the overall language stream, and moreover, more quickly influenced by verb constraints than motion constraints. In addition, each moderated the effects of the other: in the presence of agent motion, causative verbs led to faster SOTs than perception verbs, and in the presence of language, the toward condition led to faster SOTs than the neutral condition. Thus, despite the relative *primacy* of naturalistic dynamic scenes including a human figure in attracting visual fixation, it seems that this information is less *informative* than that provided by the linguistic stream, particularly verb-specific information, in the construction of event meaning.

The third experiment expanded upon these findings by exploring whether further increasing the semantic salience of the linguistic context would affect eye movement patterns. Specifically, we contrasted two versions of each initial sentence clause; one that was semantically related to the unfolding event, and one that was not. Making the initial clause more semantically restrictive should have constrained the interpretation of the unfolding event to include the participation of the target object. Therefore, upon encountering the more semantically restrictive causative verbs, saccades should have been launched more quickly to the target object. This hypothesis, however, was not confirmed: as in previous studies, verb effects only emerged when the motion context was biased toward the target object. However, constraining the initial part of the linguistic stream did affect the pattern of eye movements; in the restrictive semantic context, there were significantly more saccades, and longer fixations, to the target object prior to the verb's utterance, and saccades were launched more quickly after the onset of

the verb overall. Taken together, these results indicate that this manipulation did result in constraining the interpretation of the unfolding event.

Despite these multiple modifications to the dynamic visual world paradigm, verb effects failed to emerge across all motion contexts, only occurring when the agent's path of motion conformed to the event's meaning. Although the use of dynamic scenes is a more ecologically valid method of studying situated language processing, it does not appear to be sensitive enough to the verb's thematic properties when scenes are dynamic. These findings were consistent across all three experiments reported here, as well as in Di Nardo (2005), but are not in line with the findings of previous studies focused on the properties of verbs (Altmann & Kamide, 1999; Boland, 2005; Knoeferle & Crocker, 2007). Although the present results indicate that language comprehension mechanisms do influence gaze patterns, and further, that verb constraints do in fact produce faster saccades than agent motion (in the absence of language), it appears that the presence of human movement captures a visual attention to a greater extent when both are present. This is evidenced in two ways: one, the consistently robust main effects that motion type produced; and two, the late onset of eye movements overall.

First, the most consistent finding across all experiments and analyses, was that agent motion strongly biased the visual system toward the target object. However, these eye movements in response to agent cannot be considered simply a function of scene properties. Although saccades reached the target object before the agent did, they were locked to the agent for a short time before landing on the target object. Thus, they were not triggered by the detection of sudden movement but rather as a function of higher-order goal processes (Hillstrom & Yantis, 1994). We take this goal to be the building of

an accurate event interpretation that is primarily based on the linguistic utterance, but that also relies on agent movement. Further support for this notion comes from the study of change blindness conducted by van de Velde (2008). Using similar materials (sentences with contrasting verb types embedded within dynamic scenes), this work demonstrated *inattention* to the dissolving of the target object, which did not affect the processes of linguistic interpretation. Thus, eye movements were not affected by the sudden change in object presence but rather were locked to higher-order cognitive processes.

Second, we interpret these late saccadic onsets as reflecting a matching process between the linguistic interpretation of the unfolding utterance and the visual interpretation of the event being depicted. In other words, the construction of the event meaning relies upon integrating information from both sources, with a higher-level cognitive system (such as CSTM; Potter, 1999) seeking confirmation of the current interpretation via visual search mechanisms. Because saccades were launched most quickly in the causative-toward condition, yet occurred only after the utterance of the noun, the visual attention system appears to only make use of verb thematic information when the visual context is consistent with the meaning of the linguistic utterance. By definition, high-level central cognitive systems make use of representations generated by peripheral, domain-specific processors (Fodor, 1983). In the static scenes used in other studies (e.g., Altmann & Kamide, 1999; Boland, 2005), the representations activated by the objects populating these scenes are also static. This allows the central cognitive and attentional systems to focus primarily on the dynamically changing representations encoded by the linguistic stream, including verb-specific representations. In contrast, dynamic scenes generate object (and human entity) representations whose location and

relation to each other is constantly being updated. Thus, the higher-level systems must continually co-integrate these two main sources of information, with the highly salient visual context relegating the thematic roles encoded by verb to a lesser position of influence. Thus, in static scenes, only the linguistic stream is dynamically changing, and therefore its properties are more informative than the scene in contributing to event interpretation. On the other hand, in dynamic scenes, the position and activities of the human figure are constantly changing, and become at least as informative as the utterance in the contribution of event meaning.

The Nature of the Interaction Between Visual and Linguistic Representations

This interpretation of our results begs the question of the precise nature, and locus, of this interaction between visual and linguistic representations. While a complete account is beyond the scope of the work presented here, several theories can contribute to a better, albeit speculative, understanding of this interaction. This interaction can be framed both in terms of the nature of the representations that are integrated, and a possible model for how this occurs.

Jackendoff's (1987; see also Jackendoff, 1983) theory of conceptual semantics proposes that there are a number of primitive conceptual categories, which can include, but are not limited to, objects, events, states, or places. These combine according to certain formation rules and can be represented in the form of propositional structures (first proposed by Kintsch, 1974). The properties of a given event marker (verbs, within the language domain) restrict which agents and patients are licensed by that marker. For example, the sentence (which describes an event meaning), *The woman will crack the egg*, can be expressed in the following notation, [$_{\text{EVENT}}$ CRACK (WOMAN, EGG)],

where CRACK is the event being described, WOMAN is the entity performing the event (the *Agent*), and EGG is the entity upon which the event is being performed (the *Patient*). While this representation can be extracted from the linguistic stream, an inspection of the scene and its constituent entities can also contribute to the activation of this representation as a possible event, perhaps among many.

Importantly, these representations are conceptual in nature, and as both the utterance and the scene unfold, the selection of the appropriate event representation, or the generation of new ones, can occur via the dynamically updated information from both the visual and linguistic contexts. The features encoded by each concept (both objects and events) are based on real world knowledge, such as the typical *Agents* and *Patients* examined by Knoeferle and Crocker (2007). In the linguistic stream, verbs serve as a rich source of typical thematic roles and event predicates. However, these conceptual representations can also be activated by the entities within the scene, and dynamic scene processing likely results in the indexing of multiple representations (both entities and action goals) that take the form of visual predicates, such as those proposed by Pylyshyn (2000). These predicates are the product of situated or embodied cognition, which is precisely the type of language processing the visual world paradigm examines—and in some cases (e.g., Knoeferle & Crocker, 2007), is designed to investigate. Therefore, we propose that the outputs of both the language and visual processing systems are encoded at a common conceptual (and post-perceptual) level to produce event structures such as those proposed by Jackendoff (1987).

Given the likely conceptual nature of how events are represented, the issue then becomes focused on *how* these representations are integrated, as the overarching purpose

of the research presented here is to help inform how the interaction of the linguistic, visual and memory systems interact to construct event meaning. While much of the literature from the visual world paradigm has sought to define this interaction in terms of the modularity debate, namely whether the systems interact incrementally or at a post-modular level, part of the motivation of this work has been to reframe the interaction in terms of how the cognitive architecture is responsible for event interpretation. The combination of the sentences used, which were descriptive of events about to take place, and the dynamic scenes depicting these events, present an opportunity to examine situated language processing that is highly event-focused. Thus, while eye movement patterns do reflect the ongoing processes of language comprehension and scene processing, they are also an indication of how event interpretation takes place.

The coordinated interplay account (Crocker et al., 2010) can serve as a model for integrating the three main systems involved in situated language comprehension: memory (long-term and working memory), language processing, and visual processing. In particular, this model seems uniquely suited to examining the questions related to how event meanings are constructed as it accounts for both stored thematic role knowledge and currently active visual representations. In addition, it is a model that explains how the visual-attentional system operates in conjunction with situated language processing through the observation of its measurable output, i.e., eye movement behaviour.

The coordinated interplay account (see Figure 12 for a schematic representation of this model) stipulates that the interpretation of sentences embedded within visual scenes is continually updated through the interplay between utterance-mediated attention and scene information. Importantly, this interplay is hypothesized to occur incrementally

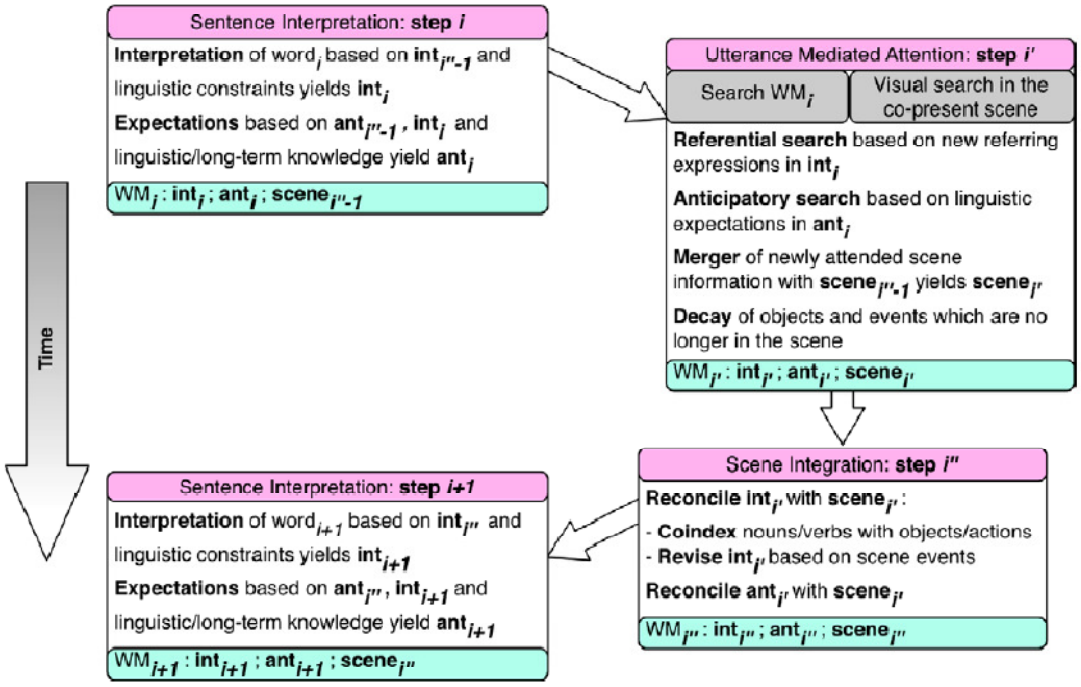


Figure 12. The coordinated interplay account (Fig. 1 in Crocker, Knoeferle, & Mayberry, 2010). The current interpretation (int_i) of the utterance is updated with information stored in long-term memory (ant_i) and salient information activated by the current scene ($scene_i$) which are incrementally integrated as the sentence unfolds.

as the sentence unfolds; as each word is uttered, the current linguistic representation limits the anticipation of possible interpretations, which leads to the referential search for the presence/interaction of entities within the scene that are consistent with that expectation. Note that this referential search is predicated upon prior processing of the scene and the activation of these entities in a working memory store. Once targeted by the attentional system, they can then be integrated and reconciled with the representation generated by the linguistic utterance. Note that although this proposal would appear to support an interactive position in the modularity debate, the precise nature of the working memory store in which these representations are proposed to interact is not described (and further, the authors state that they make no claim as to the status of the model with regards to modularity).

While the evidence upon which this model is built has employed the use of static scenes, it rather neatly accounts for the results obtained by the present work with dynamic scenes. Given the recursive nature of the coordinated interplay account, in which interpretations of both spoken language and the visual context are continually updated via the repeated sequence of processing steps (see Figure 12), the visual search triggered by sentence interpretations can not only make use of static objects, but also the dynamic motion introduced by the scene's agent. In fact, the model stipulates that scene information is highly relevant in aiding sentence interpretation. Thus, the scenes employed by the present research, which can be described as highly informative due to their realistic depiction of dynamic events, likely play a major role in situated language processing if their integration occurs in the manner proposed by Crocker and colleagues (2010).

What remains indeterminate, however, is the locus of this interaction. If, as the coordinated interplay account suggests, it occurs within working memory store, and the representations within that store are conceptual in nature, the conceptual short-term working memory (CSTM) theory proposed by Potter (1999) acts as a likely candidate for that locus. Because the CSTM draws upon conceptual knowledge stored in long-term memory, but is able to access such knowledge from the short-term activation of concepts derived from both the visual and linguistic contexts, it serves as a bridge between these two sources of information. Specifically, for the purposes of the work presented here, it allows for the integration of verb-specific thematic roles (stored in long-term knowledge) and visually present entities currently depicted in the scene, including information conveyed by their movement. Thus, the advantage conferred by the causative verbs in programming quicker saccades only when the agent in the scene is moving toward the target object can be explained in terms of this integration at a post-perceptual, conceptual stage of processing. In addition, the relatively late effect of this integration can be accounted for by the complex interplay between dynamically changing sentence *and* scene interpretations, whose computations likely consume a larger proportion of cognitive resources than when scenes are static. This likely results in the delayed effects displayed by the visual-attentional system in the experiments reported here. Given the additional goal of the central cognitive system in matching these representations to construct a higher-order event representation, as proposed by Jackendoff (1987), these late eye movements can also be seen as confirming the most probable current interpretation of the unfolding event.

Contributions, Limitations and Directions for Future Research

The work presented here serves as a significant extension on the visual world paradigm, and introduced several methodological modifications to improve upon the limitations of our previous work with dynamic scenes (Di Nardo, 2005). Most notably, we attempted to make these modifications in the service of allowing verb effects to emerge. These included eliminating the potential confound inherent in using a computer screen with a small visual angle (namely, the potential for shifts of attention without corresponding eye movements), as well as examining whether the linguistic context has any influence on gaze control, and finally, manipulating the initial semantic context conveyed by the first clause of the spoken sentence. This represents a significant contribution to the visual world literature, as no studies have yet explored how verbs might constrain the domain of visual reference in dynamic scenes, nor which variables can serve to increase the salience of these constraints.

In addition, Experiment 2 serves as a particularly important study of whether, in fact, spoken language can influence visual fixation in dynamic scenes, given the general lack of verb effects found in Experiment 1. As the visual world paradigm is predicated on the notion that features of the linguistic stream have some control of the search for visual targets, this study serves as a significant substantiation of this basic assumption. The results of this experiment clearly show that the presence of spoken language does lead to a different pattern of eye movements than that which would result from a visual inspection of the movie alone. That is, utterances exert some control over attentional processes in the referential search for targets in dynamic scenes. Furthermore, this experiment provided evidence that insofar as event meanings are predicated upon the

joint contribution of both linguistic and visual input, the relative contribution of each is in favour of the language context. Specifically, saccades to the target object were launched more quickly when the causative verb was uttered, even without corroborating evidence from the agent's motion, than when the agent moved toward the target object in the absence of spoken language. This constitutes preliminary but crucial evidence for the preference of the cognitive architecture responsible for the construction of event meaning for linguistic over visual input (although the combination of both sources of information is the true preference, as evidenced by the consistent finding that saccades are launched most quickly when both the verb and motion contexts are restrictive).

Finally, the third experiment showed that the linguistic stream can bias the interpretation of the unfolding event even without direct reference to the object involved in that event. The constraints imposed by the initial clause, which did not specifically name the target object, were sufficient to increase the number and duration of fixations to that object. Despite the inability of this manipulation to increase sensitivity to verb constraints, this builds upon the evidence of the second experiment for the importance of the linguistic stream in contributing to the construction of event meaning. This manipulation represents an important refinement to the dynamic visual world paradigm, indicating that the semantic properties of phrase segments do influence the pattern of eye movements. Furthermore, it also provides support for the notion that visual search patterns are confirmatory, rather than anticipatory, as the more semantically restrictive initial clauses did not cause saccades to be launched before the offset of the noun.

There were, however, two main limitations to the research presented here, mainly methodological and statistical in nature. First, because of the difficulties inherent in the

eye-tracker used, which was unable to prevent or correct head movements, a large proportion of data was missing (up to 30% of cell means). Despite substituting this missing data with condition means, this likely posed an issue in the validity of the analyses conducted. In addition, due to the large number of variables studied (up to four in the analyses of early post-verb saccades), the statistical power of some the analyses was reduced. Thus, the failure to find main effects of verb type in these analyses might have been due to this low power.

Second, we did not directly quantify the number and timing of saccades to the human agents. Given their role in contributing to event meaning, and given the informal observation that participants remained fixated on them for a large portion of the trials, future studies should systematically measure these eye movements. In particular, the hypothesis that human figures constitute an important source of information not just about likely *events*, (namely, the change of state in objects specified by causative verbs) but of *states*, specifies that humans are more likely to attract fixations after the utterance of perception verbs, whose thematic features emphasize an *Experiencer* role. Contrasted with the *Theme* role, which is filled by the target object, the entity undergoing the most significant change is in fact the human agent, who goes from a state of not-perceiving to perceiving. Here, the event meaning being constructed shifts from being overt to covert, and thus less depictable. Therefore, one would expect perception verbs to lead to a greater proportion of or longer fixations to the human figure than causative verbs.

In addition to the thematic roles encoded by verbs, verb tense is another syntactic variable that is highly suited to study within dynamic scenes. Altmann and Kamide (2007) have examined the role of verb tense in guiding eye movements, but with static

scenes. In depictions contrasting, for example, two glasses, one of which is full and one empty, and sentences such as *the man will drink...* or *the man has drunk...*, saccades were launched more frequently toward the appropriate visual referent (i.e., past tense—empty glass; future tense—full glass). While no such contrast was used in the studies reported here or by Di Nardo (2005), the use of the future tense in our verbs could be seen as either consistent or inconsistent with the visual context, particularly for the causative verbs. As the utterance unfolded, scene information in the form of the agent's movement was incrementally being integrated to construct the event meaning. Thus, in the away (in Experiment 1, and Di Nardo, 2005) and neutral conditions, the agent failed to move in the direction specified by the verbs (especially the causative verbs). This might have slowed the initiation of saccades to the target object as participants awaited confirmation of their interpretation of the event from the agent. Future studies could examine the role that verb tense has in influencing eye movement patterns, and how it contributes to the interpretation of events depicted by dynamic scenes.

Finally, modifications to the acoustic features of linguistic stream itself can also provide further insight into the ability of verbs to predict likely role fillers. First, the rate of speech stream could be substantially slowed down (as done in Experiment 1 of Boland, 2005). Given that the use of dynamic scenes likely occupies a greater proportion of cognitive resources, reducing the rate of speech could allow enough time for verb-driven anticipatory eye movements to occur, as the utterance of the noun would occur at a later point. Similarly, a linguistic manipulation analogous to the "blank screen" (Altmann, 2004) could be introduced such that the sentences are stopped mid-stream, just after the utterance of the verb. If the thematic roles encoded by the two verb types are in

fact able to constrain saccadic targets, then causative verbs should lead to a greater proportion of fixations to the target object after their utterance than perception verbs. This would indicate that participants are able to anticipate the likely object involved in that event, despite its not having been named.

In summary, the research presented here indicates that when realistic dynamic scenes are employed within the visual world paradigm, the linguistic information contained within the verb does not serve to constrain the domain of visual reference as found in previous studies (e.g., Altmann & Kamide, 1999; and Kamide et al., 2003; Knoeferle & Crocker, 2007), unless the agent displays movement toward the target object consistent with the utterance. Nevertheless, these studies do indicate that the linguistic context does play an important role in guiding visual attention, and in contributing to the building of event meanings. Moreover, the results presented here are consistent with the notion that the language and visual systems process information independently, and likely output this information in the form of conceptual representations that are integrated at a post-perceptual level. The coordinated interplay account (Crocker et al., 2010) serves as a promising model of how this interaction occurs, and future studies should continue to address how language processing situated in a dynamic visual world operates within the cognitive architecture. In so doing, we can develop a richer understanding of how we make use of both linguistic and visual input in building event meanings.

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,* 419-439.

Altmann, G. T. (2004). Language-mediated eye movements in the absence of a visual world: The 'blank screen paradigm.' *Cognition*, *93*, B79-87.

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

Altmann, G.T.M., & Kamide, Y. (2004). Now you see it, now you don't: Mediating the mapping between language and visual world. In J. Henderson & F. Ferreira (Eds.), *The interface of language, vision, and action: Eye movements and the visual world* (pp. 347-386). New York: Psychology Press.

Boland, J. E. (2005). Visual arguments. *Cognition*, *95*, 237-274.

Buswell, G. T. (1935). *How People Look at Pictures*. Chicago: University of Chicago Press.

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

Cooper, R.M. (1974). The control of eye fixation by the meaning of spoken language. *Cognitive Psychology, 6,* 84-107.

Crocker, M.W., Knoeferle, P., & Mayberry, M.R. (2010). Situated sentence processing: The coordinated interplay account and a neurobehavioral model. *Brain and Language*, *112*, 189-201.

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

Di Nardo, J. C. (2005). Eye movements as a function of spoken sentence comprehension and scene perception (M.A. dissertation, Concordia University, 2005). *Dissertation Abstracts International, 44*(03), AAT MR10180.

Dowty, D.R. (1991). Thematic proto-roles and argument selection. *Language, 67,* 547-619.

Dunham, P. J., Dunham, F., & Curwin, A. (1993). Joint-attentional states and lexical acquisition at 18 months. *Developmental Psychology*, 29, 827–831.

Eberhard, K.M., Spivey-Knowlton, M.J., Sedivy, J.C., & Tanenhaus, M.K. (1995). Eye movements as a window into real-time spoken language comprehension in natural contexts. *Journal of Psycholinguistic Research, 24,* 409-436.

Findlay, J.M., & Gilchrist, I.D. (2001). Visual attention: The active vision perspective. In M. Jenkin & L.R. Harris (Eds.), *Vision and Attention* (pp. 83-103). New York: Springer-Verlag.

Fischer, B., & Breitmeyer, B. (1987). Mechanisms of visual attention revealed by saccadic eye movements. *Neuropsychologia, 25,* 73-83.

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

Harris, M., Jones, D., Brookes, S., & Grant, J. (1986). Relations between the non-verbal context of maternal speech and rate of language development. *British Journal of Developmental Psychology*, 4, 261–268.

Henderson, J. M., & Ferreira, F. (2004). Scene perception for psycholinguists. In J. M. Henderson and F. Ferreira (Eds.), *The interface of language, vision, and action: Eye movements and the visual world* (pp. 1-58). New York: Psychology Press.

Hillstrom, A.P., & Yantis, S. (1994). Visual motion and attention capture. *Perception and Psychophysics, 55,* 399-411.

Intraub, H. (1999). Understanding and remembering briefly glimpsed pictures: Implications for visual scanning and memory. In V. Coltheart (Ed.), *Fleeting memories: Cognition of brief visual stimuli* (pp. 47-70). Cambridge, MA: MIT Press.

Jackendoff, R.S. (1983). *Semantics and cognition.* Cambridge, MA: MIT Press.

Jackendoff, R.S. (1987). On beyond zebra: The relation of linguistic and visual information. *Cognition, 26,* 89-114.

Kamide, Y., Altmann, G.T.M., & Haywood, S.L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language, 49,* 133-156.

Kintsch, W. *The representation of meaning in memory.* Hillsdale, NJ: Erlbaum, 1974.

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

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

Kucera, H. & Francis, W.N. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

Levin, B. (1993). *English verb classes and alternations: A preliminary investigation*. Chicago: University of Chicago Press.

Levin, B., & Rappaport Hovav., M. (2005). *Argument realization.* New York: Cambridge University Press.

Morand, S.M., Grosbras, M.-H., Caldara, R., & Harvey, M. (2010). Looking away from faces: Influence of high-level visual processes on saccade programming. *Journal of Vision, 10:16,* 1-10.

McRae, K., Ferretti, T.R., & Amyote, L. (1997). Thematic roles as verb-specific concepts. *Language and Cognitive Processes, 12,* 137-176.

Potter, M.C. (1976). Short-term conceptual memory for pictures. *Journal of Experimental Psychology: Human Learning and Memory, 2,* 50-522.

Potter, M.C. (1993). Very short-term conceptual memory. *Memory & Cognition, 21,* 156-161.

Potter, M.C. (1999). Understanding sentences and scenes: The role of conceptual short-term memory. In V. Coltheart (Ed.), *Fleeting memories: Cognition of brief visual stimuli* (pp. 11-46). Cambridge, MA: MIT Press.

Pruden, S. M., Hirsh-Pasek, K., Golinkoff, R. M., & Hennon, E. A. (2007). The birth of words: Ten-month olds learn words through perceptual salience. *Psychological Science*, 18, 414–420.

Pylyshyn, Z.W. (2000). Situating vision in the world. *Trends in Cognitive Sciences*, *4*, 197-207.

Sedivy, J.C., Tanenhaus, M.K., Chambers, C.G., & Carlson, G.N. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition, 71,* 109-147.

Spivey, M.J., Tanenhaus, M.K., Eberhard, K.M., & Sedivy, J.C. (2002). Eye movements and spoken language comprehension: Effects of visual context on syntactic ambiguity resolution. *Cognitive Psychology, 45,* 447-481.

Spivey, M.J., Tyler, M.J., Eberhard, K.M., & Tanenhaus, M.K. (2001). Linguistically mediated visual search. *Psychological Science, 12,* 282-286.

Tanenhaus, M.K., Magnuson, J.S., Dahan, D., & Chambers, C. (2000). Eye movements and lexical access in spoken-language comprehension: Evaluating a linking hypothesis between fixations and linguistic processing. *Journal of Psycholinguistic Research, 29,* 557-580.

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

van de Velde, C. (2008). Tracking eye movements to uncover the nature of visual-linguistic interaction in static and dynamic scenes (Ph.D. dissertation, Concordia University, 2008). *Dissertation Abstracts International, 69*(04), AAT NR37735.

Weith, M., Castelhano, M. S., & Henderson, J. M. (2003, May). *I see what you see: Gaze perception during scene viewing*. Presented at the annual meeting of the Vision Sciences Society, Sarasota, Florida.

Yarbus, A. L. (1967). *Eye movements and vision.* New York: Plenum Press.

Table 1

Analysis of Variance for the Effect of Sentence Point, Verb Type and Motion Type on the Cumulative Number of Saccades Initiated Towards the Target Object (by participants)

Source	*df*	*F*	η^2	*p*
Sentence Point (SP)	2	46.85	1.00	< .0001***
Verb Type (VT)	1	.35	.09	.56
Motion Type (MT)	2	4.74	.78	.01**
SP X VT	2	.06	.06	.95
SP X MT	4	6.13	.99	.0002***
VT X MT	2	2.45	.46	.09
SP X VT X MT	4	1.34	.40	.26
Error	527			
Total	544			

Notes: **$p < .01$, ***$p < .001$

Table 2

Analysis of Variance for the Effect of Sentence Point, Verb Type and Motion Type on the Cumulative Number of Saccades Initiated Towards the Target Object (by items)

Source	*df*	*F*	η^2	*p*
Sentence Point (SP)	2	68.76	1.00	< .0001***
Verb Type (VT)	1	.66	.12	.43
Motion Type (MT)	2	3.02	.54	.06
SP X VT	2	.26	.09	.78
SP X MT	4	4.96	.96	.001***
VT X MT	2	1.39	.27	.26
SP X VT X MT	4	.68	.21	.61
Error	272			
Total	289			

Note: ***$p < .001$

Table 3

Analysis of Variance for the Effect of Verb Type and Motion Type on Saccade Onset Time (by participants)

Source	*df*	*F*	η^2	*p*
Verb Type (VT)	1	.63	.12	.43
Motion Type (MT)	2	8.24	.96	.0007***
VT X MT	2	3.91	.68	.02*
Error	155			
Total	160			

Notes: *$p < .05$, ***$p < .001$

Table 4

Analysis of Variance for the Effect of Verb Type and Motion Type on Saccade Onset Time (by items)

Source	*df*	*F*	η^2	*p*
Verb Type (VT)	1	.83	.13	.38
Motion Type (MT)	2	5.28	.81	.01**
VT X MT	2	2.78	.50	.08
Error	80			
Total	85			

Note: **$p < .01$

Table 5

Analysis of Variance for the Effect of Sentence Point, Language Context, Verb Type and Motion Type on the Cumulative Number of Saccades Initiated Towards the Target Object (by participants)

Source	*df*	*F*	η^2	*p*
Sentence Point (SP)	2	36.82	1.00	< .0001***
Language Context (LC)	1	7.03	.74	.01**
Verb Type (VT)	1	3.35	.41	.07
Motion Type (MT)	1	6.53	.71	.02*
SP X LC	2	3.37	.61	.04*
SP X VT	2	.88	.19	.42
SP X MT	2	14.86	1.00	< .0001***
LC X VT	1	.04	.05	.84
LC X MT	1	3.60	.44	.07
VT X MT	1	.74	.13	.39
LC X VT X MT	1	2.50	.32	.12
SP X LC X VT	2	.51	.13	.60
SP X LC X MT	2	2.77	.52	.07
SP X VT X MT	2	3.37	.61	.04*
SP X LC X VT X MT	2	.01	.05	.99
Error	384			
Total	407			

Notes: *$p < .05$, **$p < .01$, ***$p < .001$

Table 6

Analysis of Variance for the Effect of Sentence Point, Language Context, Verb Type and Motion Type on the Cumulative Number of Saccades Initiated Towards the Target Object (by items)

Source	*df*	*F*	η^2	*p*
Sentence Point (SP)	2	45.95	1.00	< .0001***
Language Context (LC)	1	12.81	.95	.001***
Verb Type (VT)	1	.19	.07	.67
Motion Type (MT)	2	5.75	.64	.02*
SP X LC	2	7.59	.95	.001**
SP X VT	2	.71	.16	.50
SP X MT	2	9.00	.98	.0004***
LC X VT	1	.28	.08	.60
LC X MT	1	6.55	.70	.01*
VT X MT	1	.04	.05	.84
LC X VT X MT	1	.45	.10	.45
SP X LC X VT	2	.03	.05	.97
SP X LC X MT	2	1.87	.36	.16
SP X VT X MT	2	3.72	.66	.03*
SP X LC X VT X MT	2	1.96	.38	.15
Error	540			
Total	557			

Notes: *$p < .05$, **$p < .01$, ***$p < .001$

Table 7

Analysis of Variance for the Effect of Language Context, Verb Type and Motion Type on Saccade Onset Time (by participants)

Source	*df*	*F*	η^2	*p*
Language Context (LC)	1	2.30	.97	.15
Verb Type (VT)	1	.29	.22	.60
Motion Type (MT)	1	1.55	.33	.23
LC X VT	1	.06	.13	.81
LC X MT	1	2.87	.57	.11
VT X MT	1	.63	.06	.44
LC X VT X MT	1	.46	.49	.51
Error	60			
Total	67			

Table 8

Analysis of Variance for the Effect of Language Context, Verb Type and Motion Type on Saccade Onset Time (by participants)

Source	*df*	*F*	η^2	*p*
Language Context (LC)	1	7.56	.77	.01**
Verb Type (VT)	1	.22	.07	.64
Motion Type (MT)	1	3.23	.40	.08
LC X VT	1	1.97	.26	.17
LC X MT	1	6.71	.71	.01**
VT X MT	1	.93	.15	.34
LC X VT X MT	1	.14	.06	.71
Error	128			
Total	135			

Note: **$p < .01$

Table 9

Analysis of Variance for the Effect of Sentence Point, Semantic Context, Verb Type and Motion Type on the Cumulative Number of Saccades Initiated Towards the Target Object (by participants)

Source	*df*	*F*	η^2	*p*
Sentence Point (SP)	2	93.00	1.00	< .0001***
Semantic Context (SC)	1	4.50	.54	.04*
Verb Type (VT)	1	.31	.08	.58
Motion Type (MT)	1	4.20	.50	.05*
SP X SC	2	1.34	.27	.27
SP X VT	2	.08	.06	.93
SP X MT	2	3.31	.61	.04*
SC X VT	1	.03	.05	.86
SC X MT	1	1.88	.25	.18
VT X MT	1	3.80	.46	.06
SC X VT X MT	1	1.83	.25	.18
SP X SC X VT	2	3.78	.67	.03*
SP X SC X MT	2	2.30	.44	.11
SP X VT X MT	2	3.12	.58	.05*
SP X SC X VT X MT	2	9.24	.98	.0002***
Error	966			
Total	989			

Notes: *$p < .05$, ***$p < .001$

Table 10

Analysis of Variance for the Effect of Sentence Point, Semantic Context, Verb Type and Motion Type on the Cumulative Number of Saccades Initiated Towards the Target Object (by items)

Source	*df*	*F*	η^2	*p*
Sentence Point (SP)	2	100.39	1.00	< .0001***
Semantic Context (SC)	1	2.19	.27	.16
Verb Type (VT)	1	.42	.09	.52
Motion Type (MT)	1	2.29	.28	.15
SP X SC	2	.54	.13	.59
SP X VT	2	.26	.09	.77
SP X MT	2	2.86	.51	.07
SC X VT	1	.54	.10	.47
SC X MT	1	2.51	.31	.13
VT X MT	1	1.70	.22	.21
SC X VT X MT	1	.16	.07	.16
SP X SC X VT	2	1.71	.32	.20
SP X SC X MT	2	.29	.09	.75
SP X VT X MT	2	2.40	.44	.10
SP X SC X VT X MT	2	3.01	.54	.06
Error	384			
Total	407			

Note: ***$p < .001$

Table 11

Analysis of Variance for the Effect of Semantic Context, Verb Type and Motion Type on Saccade Onset Time (by participants)

Source	*df*	*F*	η^2	*p*
Semantic Context (SC)	1	4.81	.56	.03*
Verb Type (VT)	1	.99	.16	.33
Motion Type (MT)	1	11.50	.93	.002**
SC X VT	1	.24	.08	.63
SC X MT	1	15.47	.98	.0003***
VT X MT	1	8.04	.80	.007**
SC X VT X MT	1	5.64	.64	.02*
Error	301			
Total	308			

Notes: *$p < .05$, **$p < .01$, ***$p < .001$

Table 12

Analysis of Variance for the Effect of Semantic Context, Verb Type and Motion Type on Saccade Onset Time (by items)

Source	*df*	*F*	η^2	*p*
Semantic Context (SC)	1	1.65	.22	.22
Verb Type (VT)	1	.05	.06	.82
Motion Type (MT)	1	6.82	.69	.02*
SC X VT	1	.48	.10	.50
SC X MT	1	7.89	.76	.01**
VT X MT	1	6.60	.68	.02*
SC X VT X MT	1	1.37	.19	.26
Error	112			
Total	119			

Notes: *$p < .05$, **$p < .01$

Appendix A

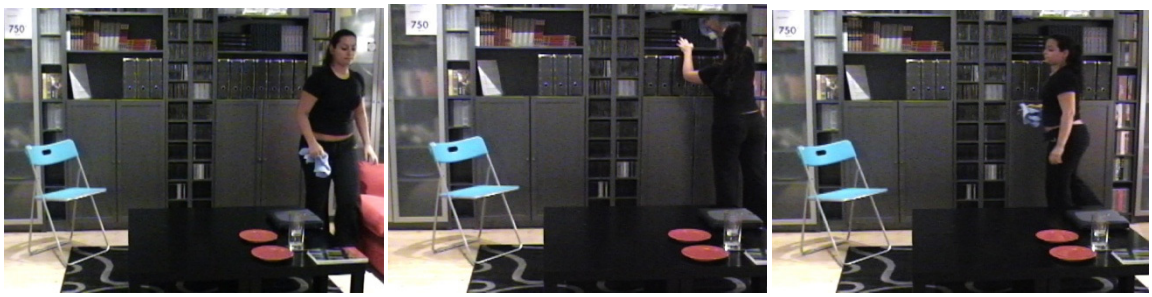
Below are the seventeen scene triplets (Away, Neutral and Towards) and the corresponding sentence pairs used in the experiment. The verb before the forwardslash (/) is the more selectionally restrictive causative verb, while the second verb is the perception verb used in each sentence pair.

1. After his warm up, the athlete will drop/inspect the ball that he uses for drills.

2. In order to bake some muffins, the woman will melt/check the butter that is required for the dough.

3. On her way to the station, the driver will crash/check the car that she just bought.

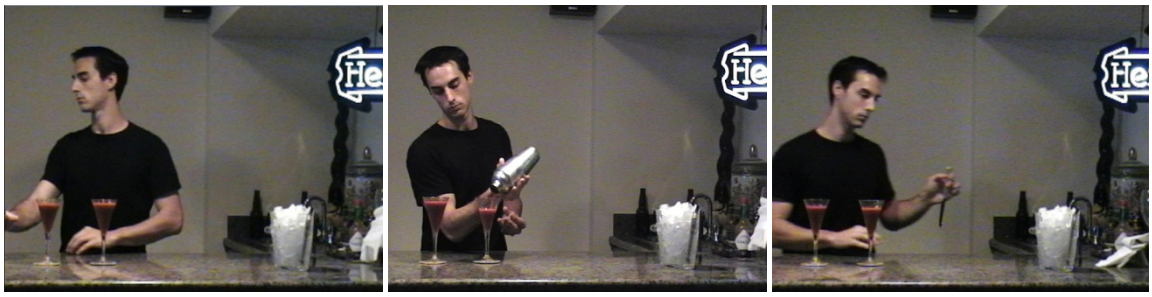
4. Before going to work, the driver will start/check the car that is in front of her house.



5. While dusting the furniture, the maid will fold/see the chair that is in the living room.

6. While playing with his toys, the infant will roll/notice the cube that is on the floor.

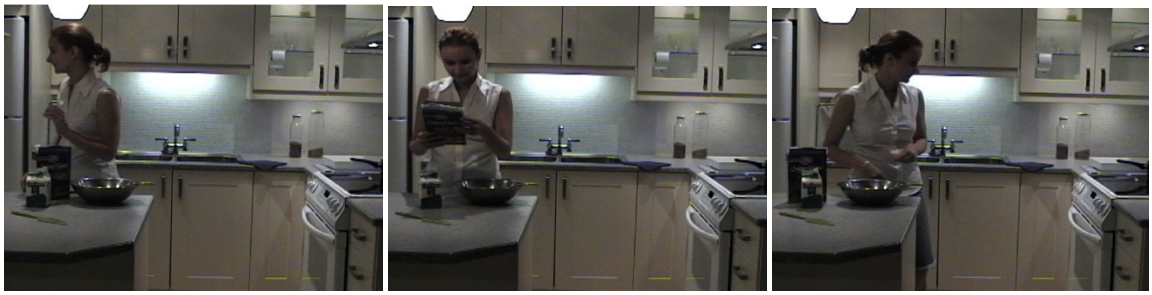
7. Before making the dessert, the cook will crack/examine the egg that is in the bowl.



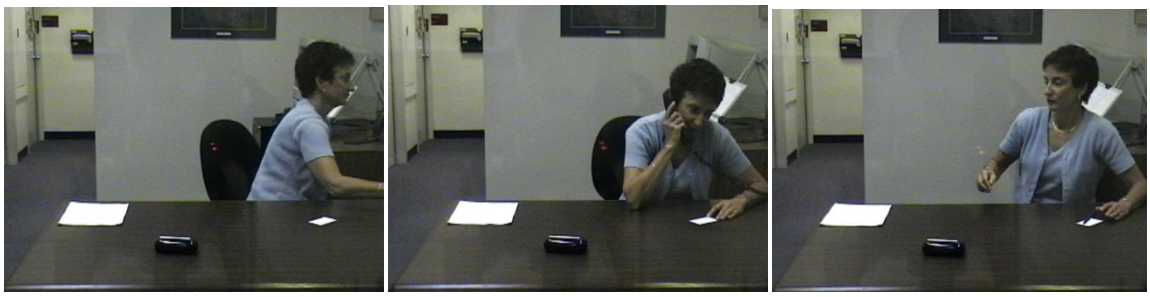
8. While preparing the drink, the bartender will crush/notice the ice that he has to put in the glass.

9. While standing in the park, the girl will fly/see the kite that is on the bench.

10. While playing with the lid, the child will spill/spot the milk that is on the table.



11. Before preparing the cake, the cook will heat/inspect the oven that is in the kitchen.



12. After talking on the phone, the secretary will rip/examine the paper that is on the desk.

13. While unpacking her office, the student will hang/study the picture that she bought at the auction.

14. Before ending his shift, the busboy will dry/spot the plate that is on the counter.

15. While packing his clothes, the man will wrinkle/see the shirt that he will use at the meeting.

16. After getting ready for work, the businessman will shine/examine the shoes that he got from his wife.

17. During her visit to the gallery, the girl will break/spot the vase that is on display.

Appendix B

INSTRUCTIONS

In this experiment, you will see a series of short movies displayed on the screen. At the same time, you will hear a sentence that refers to the event occurring on the screen. Your task will be to view the sentence-related events.

During this experiment, we will also be recording your eye movements. This will be done through the use of an eye-tracking machine as you watch the movies. You will rest your chin and forehead against the eye-tracker. This equipment does not pose any risks, although it may be slightly uncomfortable. Before the experiment begins, please inform the experimenter if you are uncomfortable so that it can be adjusted.

There are a few details to understand before starting. Please read the sequence of tasks carefully, and make sure you understand what you should do in each part of the experimental trials.

1. First, the instructions will appear on the screen. Take the time to read these carefully and ask the experimenter if you have any questions.

2. Each trial will begin with the presentation of a fixation cross (+) displayed in the middle of the screen. You should focus on this cross until it disappears.

3. When the movie begins and the fixation cross disappears, you are free to move your eyes and scan the scene.

4. It is important that you pay attention to both the visual display and the sentence presented over the earphones.

5. When the trial is over you will see a black screen for a few seconds, and then another cross will appear. This is the beginning of the next trial.

6. If you have any questions or concerns, do not hesitate to speak to the experimenter.

Have fun!

Appendix C

Thank you for choosing to participate in this experiment.

You will be presented with a series of pictures accompanied by spoken sentences relating to the scene on-screen. You are asked to simply look at the pictures and listen to the sentences. Prior to each trial, there will be a fixation cross (+) in the middle of the screen that you must fixate on. This cross will be red on a black background. The picture will then appear, with the cross still in the centre of the screen. Keep looking at the cross. Once the + disappears, you may look wherever you like on the screen. To move on to the next trial, just press the spacebar. It is important to remember to pay attention to both the pictures and the spoken sentences. After the experiment is finished, you will be given a short memory task to ensure that you have been paying attention.

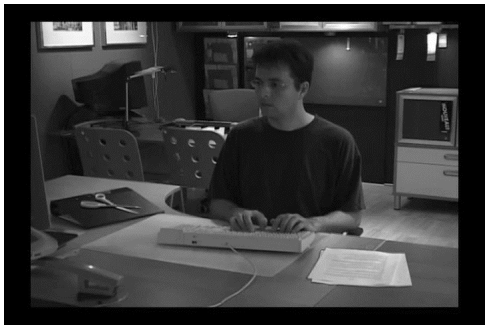
If at any time you experience discomfort, you may choose to discontinue the experiment.

Now sit back, relax, and enjoy!

Appendix D

Below are the six sentences and screenshots of the distractor trials used in Experiments 2 and 3.

1. The man is preparing an elaborate dinner to surprise his girlfriend for her birthday.



2. Because he procrastinated, the student is staying up late writing a term paper that is due tomorrow.

3. To prepare for her class, the teacher is writing up an assignment to give to her students.

4. The young man is putting together the materials he needs to build a backyard shed.

5. The girl is getting ready for a day of fun at her cottage on the lake.

6. The man is preparing his breakfast before going to work so that he won't be hungry before lunch.

Appendix E

Below is a list of the sentence quartets used in Experiment 3. Each initial patch clause had two variants, "non-restrictive" (before the forwardslash) and "restrictive" (after the forwardslash). The main and second patch clauses remained identical, as did the verb pairs.

1. After his warm-up/Before playing a game of soccer, the man will inspect/drop the ball that he will use for his match.

2. After cutting open the bag of sugar/To prevent the muffins from sticking, the woman will check/melt the butter needed to grease the pan.

3. After talking to her boyfriend/While going on a test drive, the girl will check/crash the car that she wants to buy.

4. After leaving the house/After unlocking the door, the woman will check/start the car that is parked on the street.

5. While tidying up the house/Before sitting down for a rest, the maid will see/fold the chair that is in the living room.

6. Before going to take a nap/After losing interest in the toy train, the toddler will notice the cube that is on the floor.

7. After pouring the flour into the bowl/In order to make the omelette, the cook will examine/crack the eggs that are on the counter.

8. While entertaining his date/In order to cool the drinks, the man will notice/crush the ice that is in the bucket.

9. While spending a day at the park/Because it is a windy day, the girl will see/fly the kite that is on the bench.

10. While getting ready for bed/After playing with the cup's lid, the boy will spot/spill the milk that is on the table.

11. After reading the recipe/Before baking the cake, the woman will inspect/heat the oven that is in the kitchen.

12. After her phone call/Before recycling it, the secretary will examine/rip the paper that is on the desk.

13. Before sitting down at her computer/While decorating her new office, the worker will study the picture that is on the desk.

14. Before completing his list of tasks/After washing the dishes, the man will spot/dry the plate that is in the rack.

15. Before leaving on his trip/While getting dressed for his trip, the businessman will see/wrinkle the shirt that is on the hanger.

16. After putting on his tie/After putting on his clothes, the man will examine/shine the shoes that are on the bench.

17. During her visit to the gallery/While moving on to the next piece of art, the woman will spot/break the vase that is on display.

Appendix F

INSTRUCTIONS

In this experiment, you will see a series of short movies displayed on the screen. Your task will be to simply watch these short clips.

During this experiment, we will also be recording your eye movements. This will be done through the use of an eye-tracking machine as you watch the movies. You will rest your chin and forehead against the eye-tracker. This equipment does not pose any risks, although it may be slightly uncomfortable. Before the experiment begins, please inform the experimenter if you are uncomfortable so that it can be adjusted.

There are a few details to understand before starting. Please read the sequence of tasks carefully, and make sure you understand what you should do in each part of the experimental trials.

7. First, the instructions will appear on the screen. Take the time to read these carefully and ask the experimenter if you have any questions.

8. Each trial will begin with the presentation of a fixation cross (+) displayed in the middle of the screen. You should focus on this cross until it disappears.

9. When the movie begins and the fixation cross disappears, you are free to move your eyes and scan the scene.

10. It is important that you pay attention to the visual display as your memory will later be tested.

11. When the trial is over you will see a black screen for a few seconds, and then another cross will appear. This is the beginning of the next trial.

12. If you have any questions or concerns, do not hesitate to speak to the experimenter.

Have fun!

Appendix G

Thank you for choosing to participate in this study.

- You will see a series of short movie clips; you will be asked to answer a few questions about each.
- The first TWO are practice trials to help you familiarize yourself with the task.
- Each movie begins with a fixation cross - focus on it until the movie appears.
- Each film clip will end with the last frame still showing - you may use it to help you answer the questions.
- Once you have finished answering the questions, turn the page and then press the right arrow button (→) ONCE to start the next movie.
- Please note that some of the questions may sound a bit “odd” - do your best to answer them, but don’t worry about finding a “right” answer. Also, please try to answer the questions as quickly as possible - we want your “gut reaction.” Finally, for the first three questions, please make your sentences as complete as possible.