* Manuscript

The influence of recent scene events on spoken comprehension: evidence from eye movements

Pia Knoeferle (Computational Linguistics, Saarland University, Germany)

Matthew W. Crocker (Computational Linguistics, Saarland University, Germany)

Correspondence concerning this article should be addressed to:
Dr. Pia Knoeferle
Department of Computational Linguistics
Building C71, Room 1.20
Postbox 15 11 50, Saarland University
66041 Saarbrücken
Germany

email: knoeferle@coli.uni-sb.de
telephone: +49(0)681 302 6557
fax: +49 (0)681 302 6561

Abstract

Evidence from recent experiments that monitored attention in clipart scenes during spoken comprehension suggests that people preferably rely on non-stereotypical depicted events over stereotypical thematic knowledge for incremental interpretation. The Coordinated Interplay Account (Knoeferle & Crocker, 2006) accounts for this preference through referential processing (e.g., the verb mediates a depicted event) and the preferred use of scene event information that is associated with the referent (e.g., the agent of the depicted event). Three eye-tracking experiments examined the generality of this account. While the rapid use of depicted events was replicated in all three studies, the *preference* to rely on them was modulated by the decay of events that were no longer co-present. Our findings motivate the extension of the Coordinated Interplay Account with an explicit working memory. The coordinated interplay mechanism together with working memory and decay, is shown to account for the influence of scene-derived versus stored knowledge both when events are co-present and when they have recently been perceived.

Keywords: coordinated interplay account, priority of depicted events, language-vision interaction, situated comprehension, eye tracking

Introduction

Previous research has examined the time-course with which both linguistic knowledge and scene information are applied during spoken comprehension by monitoring people's eye movements to objects in a visual context as they listened to a related utterance (e.g., Altmann & Kamide, 1999; Chambers, Tanenhaus, & Magnuson, 2004; Kaiser & Trueswell, 2005; Knoeferle & Crocker, 2006; Sedivy, Tanenhaus, Chambers, & Carlson, 1999; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995).

Findings by Tanenhaus et al. (1995), for instance, have revealed that people rapidly process word meaning, inspecting appropriate referential candidates in a scene as they are mentioned. Long-term linguistic and world knowledge can even direct people's attention to as-yet unmentioned objects (Altmann & Kamide, 1999) and role fillers in a visual context (e.g., Kamide, Scheepers, & Altmann, 2003). In turn, there is evidence that the scene itself in the form of either a visual referential context (Tanenhaus et al., 1995) or depicted clipart events (Knoeferle, Crocker, Scheepers, & Pickering, 2005) can rapidly influence the incremental structuring and interpretation of an utterance. The influence of depicted events on comprehension has also been shown to be closely temporally coordinated with when a verb identifies a relevant event in the visual context (Knoeferle, in press; Knoeferle & Crocker, 2006).

To further investigate the influence of scene information on sentence comprehension, Knoeferle and Crocker (2006, Exp. 2) examined how important depicted events are for thematic role assignment relative to stereotypical thematic role knowledge. As expected based on findings by Kamide et al. (2003) and Knoeferle et al. (2005), eye movements in clipart scenes confirmed that people rapidly exploited verb-based stereotypical knowledge of likely agents as well as depicted agent-action events in anticipating either a stereotypical agent or the agent of a verb-mediated depicted event, respectively, shortly after the verb identified one of them as relevant. In a further condition, the verb identified both a stereotypical agent and an agent depicted as performing a (non-stereotypical) event. The comprehension system was thus forced to choose between two available, yet conflicting, types of information in online thematic role-assignment. Gaze patterns suggested a clear preference of the comprehension system to rapidly rely on depicted events over stored thematic knowledge: People anticipated the agent of the depicted event rather than the stereotypical agent when both of these were identified as relevant by the utterance.

Based on the above eye-tracking findings, and further motivated by insights concerning the importance of both joint attention and scene information in language acquisition (e.g., Harris, Jones, Brookes, & Grant, 1986; Dunham, Dunham, & Curwin, 1993; Snow, 1977, see also Richardson & Dale, 2005 and Richardson, Dale & Kirkham, in press, for related evidence on joint attention in adults), Knoeferle & Crocker outlined the Coordinated Interplay Account of scene-sentence interaction in adult utterance comprehension. In its initial outline, the Coordinated Interplay Account identified two basic steps in situated comprehension. First, comprehension of the unfolding utterance guides attention in the scene, establishing reference to objects (e.g., Tanenhaus et al., 1995) and events (e.g., Knoeferle et al., 2005), and anticipating likely referents (e.g., Altmann & Kamide, 1999). Once attention has shifted to the most relevant object or event, the attended scene information rapidly influences utterance comprehension (Knoeferle & Crocker, 2006; Knoeferle, in press). The Coordinated Interplay Account further suggests that the close time-lock between comprehension and attention in the scene is at the origin of the relative priority of immediately depicted events over knowledge of stereotypical events in comprehension.

To more concretely illustrate the relative order in which utterance comprehension mechanisms interact with attention to objects and events in the scene as well as with information provided by those objects and events, we first detail the mechanism for one word in the utterance. We then show that this more detailed outline of the Coordinated Interplay accommodates the findings that originally motivated the development of the account (Knoeferle & Crocker, 2006, Experiment 2). Finally, we identify open issues regarding the Coordinated Interplay Account, which motivate the studies presented in this paper.

The Original Coordinated Interplay Account

Consider first the coordinated interplay of utterance and scene information as people encounter the i^{th} word in the utterance. For simplicity's sake, we outline key processes at word i in a quasi-serial order. However, we embrace the possibility that they may partially overlap and occur in parallel, modulo informational dependencies. At word i, the comprehension system begins to build a structure and interpretation for the utterance, deriving linguistic expectations in the process. The listener then inspects the scene for objects/events that the current input refers to, and may further explore the scene, anticipating objects/events based on his linguistic expectations. Upon finding an object/event that is the referent for the current input, reference to that object/event is established by forging a link between the meaning of the referring expression in the input and the designated object/event in the scene/world. In the model, this is achieved through coindexing of referring expressions in the utterance (e.g., nouns and verbs, see Knoeferle et al., 2005) with corresponding scene objects and events (see also Jackendoff, 2002). Proximal scene-based information is further available for potentially revising the existing interpretation of the utterance, and associated linguistic expectations. After revisions, the next word (*i+1*) is integrated with the revised interpretation, yielding a new interpretation and derived linguistic expectations.

Processes of searching for a referent, referential processing, and the use of visually associated scene information to inform utterance comprehension thus constitute the central mechanism of the model. This mechanism has further been implemented in a connectionist network model (Mayberry, Crocker, & Knoeferle, 2006). Naturally, however, not every object mentioned in the utterance will be in the scene, and often utterances will not - or will only tangentially - relate to the immediate scene/environment. The Coordinated Interplay Account assumes that the relative importance of scene information is a function of referential success throughout the processing of the utterance (see Knoeferle & Crocker, 2006, for discussion). The connectionist implementation of the account allows for these considerations: While it accounts for the use and greater relative importance of depicted events when scene information is available, it also successfully models comprehension when only linguistic input is available to the network.

To more concretely illustrate the Coordinated Interplay Account, we consider an example sentence from the second experiment reported by Knoeferle and Crocker (2006): *Den Piloten bespitzelt der Detektiv* (‘The pilot$_{acc/obj}$ spies-on the detective$_{nom/subj}$’). The corresponding clipart scene showed a pilot, a wizard spying on the pilot, and a detective serving food to the pilot. When *Den Piloten* (‘The pilot$_{obj}$’) is encountered, the meaning of *Pilot* is accessed and integrated with linguistic constraints (e.g., accusative case marking on the determiner *Den* of the noun phrase). People begin to build an interpretation that contains the pilot as the patient of some, as-yet-unknown, action. Eye movements are directed towards the pilot, and the comprehension system establishes reference to the pilot in the scene. People may further explore proximal scene regions, and notice that the pilot is the patient of two events: a wizard that is spying on the pilot, and a detective that is serving food to the pilot.

When the ensuing verb *bespitzelt* (‘spies-on’) is processed, its meaning is accessed, and the interpretation that is built consists of a spying action of which the pilot is the patient. Further, expectations about the likely agent of the spying action are informed by two sources: Verb meaning alone encourages linguistic expectations of a detective as a stereotypical agent of ‘spies-on’, while the scene provides information that the wizard is depicted as the agent of a spying action (to the extent this has been perceived). Verb-mediated processes seek a referent for the verb, initiating a visual search for a depicted spying event. Reference is subsequently established between the verb and the depicted action of the spying event. Some attention also goes to proximal objects such as the agent of the depicted spying event (the wizard). Informed by the coindexed scene event (wizard-spying) the wizard is now established as the most relevant agent for the spying action.

On the post-verbal adverb (*gleich*, 'soon'), people access its meaning, and extend the interpretation containing a pilot being spied-upon by a wizard (based on the scene depiction), with the future tense of the adverb, possibly establishing a slight expectation of a future or ongoing spying event. Given that the depicted events are not yet complete, this is consistent with the interpretation that has been built. Eye movements are now firmly directed towards the wizard rather than to the detective in the scene. Then the second noun phrase *der Detektiv* ('the detective$_{subj}$') is encountered, its meaning accessed, and the comprehension system now favours an interpretation containing a pilot that is being spied-upon by a detective. As a result, the coindexing between the verb ('spies-on') and the depicted spying action must be removed. Most eye movements now center on the detective in the scene, and people establish reference to that agent.

Open Issues: co-presence and plausibility of depicted events

In the original account, the preferred reliance on depicted events in anticipating likely role fillers is thus attributed to prioritizing utterance-mediated search for, and attention to, a referent in the scene, combined with referential processing through coindexing and an active perceptual exploration of proximal role fillers in the visual context.

The generality of the priority of depicted events, however, still needs to be established. One central issue concerns the visual co-presence of depicted events during utterance comprehension. A narrow interpretation of the original Coordinated Interplay Account implies that the visual co-presence of depicted events and their availability for inspection during online comprehension are pre-requisite for the greater relative importance of depicted events. Alternatively, depicted events that have already been perceived and that are no longer co-present during utterance comprehension may still influence comprehension online. If this were the case, then the Coordinated Interplay Account would require an explicit mechanism for exploiting information about previously-inspected objects/events.

Previous research has investigated how people attempt to access information about objects that are no longer immediately visible. Spivey and Geng (2001, Exp. 2) used scenes containing four differently coloured shapes that were either tilted leftward or rightwards. After participants had inspected the objects, one of them disappeared. When answering a question about either colour and tilt of the disappeared object, participants re-fixated its previous location on half of the trials. To perform this task, people had to access information about the colour or tilt properties of an object. The standard view is that such object properties can be accessed from internal working memory representations (e.g., Baddeley, 1986). More recently however, it has been suggested that the external world may also function as a memory, and that an object's properties are associated with its location in the external world: Working memory representations of an object thus contain a "pointer" to the object's location in the world, as well as a label for the object (e.g., 'the wizard'). The object properties are accessed through the object's label and its pointer that directs perceptual attention to the location of the object. Eye movements to the previous location of absent objects in answering questions about object properties (Spivey & Geng, 2001) were taken as support for this view (e.g., Richardson & Spivey, 2000; Spivey et al., 2004; see also Ballard, Hayhoe, Pook, & Rao, 1997).

Altmann (2004) applied Spivey and Geng (2001)'s method to the investigation of utterance comprehension. People inspected an image with a man, a woman, a cake, and a newspaper before the screen went blank ('blank-screen' paradigm). Two seconds later, people heard a sentence that described part of the previously-inspected scene (e.g., *The man will eat the cake*). Having heard the verb, people rapidly looked at the location where previously there had been a cake. The time-course of eye movements to the cake closely resembled the time-course of inspections in an earlier study with similar materials using concurrent rather than serial scene-utterance presentation (Altmann & Kamide, 1999). The data by Altmann (2004) support the external-memory account by Spivey, Richardson, and Fitneva (2004), and suggest that even when a prior scene is no longer immediately available, the mental representations that are derived from it may remain accessible for comprehension.

However, it is unclear whether findings by Altmann (2004) extend to complex scenes that contain agent-action-patient events in addition to objects and characters, and to scene information that is non-stereotypical, or even implausible, as with the depicted events in the studies by Knoeferle and Crocker (2006). Moreover, even if we find that people rapidly *exploit* depicted event information when those events are both non-stereotypical and no longer visually co-present, it is unclear whether they still *prioritize* them over long-term knowledge of stereotypical agents for thematic interpretation, as predicted by the Coordinated Interplay Account.

To address these issues, and further refine the Coordinated Interplay Account, Experiment 1 of the present paper used the stimuli from Knoeferle and Crocker (2006, Exp. 2) but changed image presentation: People inspected a scene (Fig. 1a), then the screen went blank, and people's eye movements were monitored in the blank screen during utterance comprehension. Replicating the finding that people rapidly exploit depicted event information with this presentation would show that findings from the experiment by Altmann (2004) generalize to depicted events and to scenes that depict non-stereotypical role relations. It would further entail that we extend the Coordinated Interplay Account with a working memory mechanism such as the one proposed by Spivey et al. (2004). This would enable comprehension, in addition to searching the scene, to search for referents also in working memory representations of the visual environment, using pointers for exploiting information about objects/events that are no longer co-present. Beyond the *use* of depicted events, an open issue is whether people still *prefer* to rely on them even when they are both non-stereotypical and no longer co-present. Recall that a strict reading of the Coordinated Interplay Account suggests that the visual co-presence of the depicted events is responsible for their priority. Failure to replicate the preferred reliance on depicted events from Experiment 2 by Knoeferle and Crocker would support this strict reading. Replication of the depicted events preference, in contrast, would suggest that the visual co-presence is not a precondition for the priority of depicted events, and that they are still preferred for thematic interpretation even if they have to be accessed from working memory.

Experiment 2 further investigated the role that the co-presence of stereotypical agents versus non-stereotypical depicted events plays for their respective influence during online language comprehension. One possible criticism of using the blank screen method to investigate the relative priority of events is that it disadvantages the stereotypical agent just as much as the depicted events. In real life, scene events may be at a greater disadvantage than entities regarding their co-presence. Real-world actions are often fleeting, completed, and part of the recent past, and may as a result only be briefly available for inspection, and not necessarily concurrently with utterance comprehension. Event participants, however, often remain for some time after performing an action. Furthermore, the fact that there are no agents in the blank screen for Experiment 1 that could, in principle, perform a future action may discourage expectations of a (non-depicted) stereotypical event at the verb, and encourage reliance on a past, depicted event and its associated agent for thematic interpretation.

To put the depicted event preference to a more rigorous test, Experiment 2 used sequences of four images to briefly depict the actions, then remove them with only the characters remaining as people listened to the utterance. This manner of presentation is a first - admittedly coarse-grained - approximation of more dynamic action sequences in which events are depicted as past and completed while the agents remain co-present and are available to perform further actions. If we replicate the rapid use of depicted events with such presentation of the stimuli from Knoeferle and Crocker (2006, Exp. 2), then this would suggest that the completed aspect of past depicted events does not, in principle, preclude their rapid use. Such a finding would provide further support for integrating a working memory mechanism into the Coordinated Interplay Account.

Should we, in addition, replicate the *preferred* reliance on non-stereotypical depicted events in Experiment 2, then this would suggest that even when apparently completed depicted events have to be accessed from working memory while stereotypical agents are co-present in the scene, the greater relative importance of depicted events for thematic role assignment remains. If, in contrast, we fail to replicate the greater relative priority of depicted events, then we can assume that accessing depicted events from working memory while relevant stereotypical agents are available for visual inspection, weakens the greater relative priority of depicted events. In terms of the Coordinated Interplay Account, failing to replicate the priority of depicted events would advocate that we revise the original mechanism such that representations only available in working memory are less salient than representations that still receive support from corresponding co-present objects/events. A more suitable, revised mechanism might, for instance, assume that entities and events in working memory, which are no longer in the scene, decay.

The first two experiments investigated whether the absence of depicted event scenes (Experiment 1) or their presentation as completed (Experiment 2) modulates the influence of depicted events when they are non-stereotypical. Experiment 3 addresses the same question as Experiment 2, but with new, plausible, stimuli. Just as in Experiment 2, the events are presented as completed and in the past, and thus may experience some decay, while potential role fillers are co-present in the scene. However, unlike in Experiments 1 and 2, the role filler associated with the recently depicted event is now supported by long-term knowledge derived from the verb in addition to event depiction. A further issue that we addressed in Experiment 3 was the way in which depicted events were identified by the utterance. With the more dynamic presentation, people may perceive previously depicted events as “just completed”. To felicitously identify a past event compared with an event that has not yet occurred, it may be important that the tense of the verb matches the past versus future status of the event. There is indeed some evidence that people exploit tense cues in interpreting spoken sentences with respect to a visual context (e.g. Kamide et al., 2003; Knoeferle et al., 2005; Knoeferle & Crocker, 2006). Scenes in these previous studies, however, were static. In Experiment 3, we present plausible events in a dynamic manner similar to Experiment 2, and more fully investigate the time-course with which tense cues in the utterance (e.g., verb tense and past versus future adverbs) can bias people during comprehension to access either a past depicted event or anticipate a future event that has not yet taken place.

Experiment 1

Experiment 1 investigated whether the rapid use and priority of depicted events was maintained when the scene, and its events were no longer co-present during utterance comprehension. To this end, we used the materials from Experiment 2 by Knoeferle and Crocker (2006). Unlike in their experiment, however, people inspected the clipart scene prior to rather than during comprehension (e.g., Fig. 1a).

After inspection and prior to utterance presentation, the scene was replaced by a blank screen.

Method

Participants

Fourty native speakers of German with normal or corrected-to-normal vision received five euros each for participating in the experiment.

Materials and Design

The materials, design, and counter-balancing were the same as in Knoeferle and Crocker (2006, Exp. 2). Scenes, just as in their studies, thus were *Ersatz scenes* rather than real-world environments (see Henderson & Ferreira, 2004, for a discussion on different types of scenes). The scenes in our experiments were created from clipart images with the help of commercially available graphic packages. There were 24 items and four experimental conditions (Table 1, (a1), (a2), (b1), (b2), and Fig. 1a). For counter-balancing reasons, however, one item comprised eight sentences and two images. We describe the four experimental conditions, and then the counter-balancing.

Figure 1: *about here #*

As in the original study, participants in Experiment 1 heard unambiguous German object-verb-subject (OVS) sentences. While the OVS structure is non-canonical in German, this was a constant across conditions. We used item sentences with an OVS order for design-related reasons of avoiding visual confounds, and mediating relevant events early through the verb. Item sentences in subject-verb-object order that relate to agent-action-patient events, for instance, would have resulted in depiction biases: If the pilot had been the subject and agent of the sentence, he would have been acting upon and facing either one of the other two characters, creating a visual bias. The use of passive sentences in which the pilot is a patient might have avoided such depiction biases. Passive sentences in German, however, are verb-final, and thus would not have enabled us to test the early, verb-mediated use of depicted events versus verb-based thematic role knowledge.

Sentences started with an accusative noun phrase (*Den Piloten*$_{acc/obj}$..., 'The pilot$_{acc/obj}$ ...') identifying a character that was the patient of two depicted agent-action events in the previous scene (the pilot, Fig. 1a, Table 1). The depicted events provided information about role relations (e.g., wizard-spying-on-pilot and detective-serving-food-to-pilot, see Fig. 1a). In addition, each agent provided stereotypical thematic role knowledge (e.g., a detective is a stereotypical agent of a spying, and a wizard a stereotypical agent of a jinxing action; characters were never depicted doing that stereotypical action in the experimental items).

In a first condition pair, the utterance continued either with the verb *verzaubert* ('jinxes'), or with the verb *verköstigt* ('serves-food-to') (Table 1). Each of these verbs uniquely identified one agent in the previous scene by means of either knowledge about stereotypical agents of a non-depicted jinxing action (wizard, (a2)), or through a non-stereotypical event but that was depicted in the previous scene (detective-serving-food, (a1)). If the co-presence of the scene is not essential for the reliance on stereotypical thematic role knowledge and non-stereotypical depicted events, people should rapidly exploit either stereotypical knowledge of likely agents ('jinxes') or depicted events (serving-food) for incremental thematic interpretation at the verb. This should be revealed by more anticipatory inspection shortly after the verb to the region in the blank screen where the appropriate stereotypical agent (the wizard) or the agent of a depicted food-serving event (the detective) had previously been.

In a further condition pair, people heard 'The pilot (object) spies-on ...'. In this case, the previously-inspected scene afforded both a stereotypical agent (the detective), and the agent of a depicted event (wizard-spying) as relevant agents of the verb (see Fig 1a). Thus, stereotypical knowledge of 'spies-on' and scene events identified both of these agents. The comprehension system had to choose between two available, yet conflicting, types of information in online thematic role-assignment. If people prefer to rely on non-stereotypical depicted events even in the blank screen, they should anticipate the region of the agent depicted as performing the action (the wizard in Fig. 1) more often than the region where the stereotypical agent (the detective) had been, shortly after hearing the verb.

Table 1: *about here #*

Manipulation of the verb created four conditions, crossing the factors *target type* (depicted, stereotypical) with *identification* (unique, ambiguous). 'Identification' refers to whether the verb uniquely identifies a target agent based on either depicted events or stored thematic knowledge ((a1) and (a2) respectively), or whether it identifies both relevant available agents (a stereotypical agent and the (different) agent of a depicted action) as relevant ((b1) and (b2)) ('ambiguous identification'). 'Target type' refers to which target agent type (depicted, stereotypical) the second noun phrase identifies as the appropriate agent. In conditions (a1) and (b1) the target type is depicted and in conditions (a2) and (b2) the target type is stereotypical. Note the difference between 'target type' and 'identification': 'Identification' characterizes which informational source is identified as relevant by the *verb*, whereas ‘target type’ refers to the target agent that is identified by the *second noun phrase*. When referring to the target agent identified by the second noun phrase we will use the expression ‘target type’. When we refer to the scene entities, we will use ‘target agent’ (depicted agent and stereotypical agent).

Materials were designed so that any plausibility biases introduced by the verb or noun phrases, depiction biases introduced by characters, or position biases (whether an agent was to the left or right of the patient) were counter-balanced. We created two images for an item (Fig. 1) to ensure that each verb once identified the agent of a depicted event as relevant (e.g., *verzaubern*, ‘jinx’, detective-jinxing; Table 1, (b1’), and Fig. 1b), and once a stereotypical agent (‘jinx’, wizard-jinxing; Table 1 (a2), and Fig. 1a). It further ensured that each of the agents (wizard, detective) was once identified as a potential agent through verb-mediated depicted actions (e.g., wizard-bandaging; Table 1, (a1’), and Fig. 1b) and through stereotypical verb-based knowledge (e.g., a wizard is a stereotypical agent of a jinxing action, Table 1, (a2), and Fig. 1a). Each agent was once uniquely (e.g., the detective in (a1)) and once ambiguously (e.g., the detective in (b2)) identified as relevant through the utterance. As a result of the counter-balancing, conditions were matched for length (number of spoken syllables), and frequency of lemmas within an item (t-test, p = 1) (Baayen, Pipenbrock, & Gulikers, 1995). The mean length (number of spoken syllables) and mean log frequency were identical across conditions for the first noun (mean length: 2.63, mean log frequency: 0.72), the verb (mean length: 2.23, mean log frequency, 0.70), the adverb (mean length: 1.83, mean log frequency: 1.18), and the second noun (mean length: 2.94, mean log frequency: 0.39). Off-line plausibility ratings ensured that the character which was the stereotypical agent for a verb (detective, ‘spy-on’) was a more plausible agent for that verb than the other character (wizard, Fig. 1) (see Knoeferle & Crocker, 2006).

In addition to the 24 items, there were 48 filler items. The filler items were balanced for the number of stereotypical versus depicted actions and agents. They further contained twelve filler trials that showed an agent performing a stereotypical action. This was done to minimize the possibility that participants pay less attention to their stored linguistic and world knowledge than to the immediately (implausible) depicted events in on-line comprehension. Such a strategy could be induced if participants never encountered a plausible action that was also depicted. Including twelve filler items where a plausible action was depicted should help to avoid this problem. Among the filler items, 40 were subject-initial sentences to ensure that the majority of sentences a participant heard were canonical subject-initial sentences. Experimental items were separated from one another by at least one filler item. There were eight experimental lists with 72 trials. Each of the 40 participants heard only one of the eight sentences of an item, and the order of appearance of items was randomized individually for every participant.

Procedure

An SMI Eye-Link head-mounted eye-tracker monitored participants' eye movements at a frequency of 250 Hz. Images were presented on a 21-inch multi-scan color monitor at a resolution of 1024 x 768 pixel. Participants were seated at a distance of approximately 50 centimetres from the monitor. The width of a typical scene object was approximately 5 centimetres, and the corresponding visual angle with center vision was 6 degrees. Before the experiment, participants received written instructions that were similar to the instructions in the blank screen study by Altmann (2004): Participants were asked to inspect the images attentively, and informed that the study examined what happened when people listen to sentences that describe events that either were depicted on a previously viewed scene, or that might take place. Note that for the unique stereotypical target conditions the verb refers to a non-depicted action in the future (e.g., 'jinx', a2, Table 1), and the second noun phrase identifies the wizard as the agent despite the fact that the wizard is depicted as performing another action (spying). Unless people pay attention to verb and adverb tense (future), they may perceive a potential clash between the jinxing event described in the utterance and the fact that the wizard is not currently depicted as jinxing. However, we expected that in light of the instructions, a description of a future event would not appear too clashing. Furthermore, people did not make the experimenter aware of any observed mismatches. Note also that at our critical region of measurement (the adverb region), people had not yet heard the entire sentence, and thus could not be aware of the potential clash between a sentence such as 'The pilot (obj) will soon be jinxed by the wizard (subj)' and the events depicting the wizard as spying. Thus, such a potential clash – even if it existed – could not influence the anticipatory eye movements during the adverb region. For each trial, the image was presented first, and participants inspected it for a fixed duration of 7000 ms. The image then disappeared, and the screen went blank. After a delay of 1500 ms, the utterance was presented. The entire experiment lasted approximately 30 min.

Analysis

The eye-tracker software recorded the X-Y co-ordinates of participants' fixations. To analyse the output of the eye-tracker, the visual scenes were coded into distinct regions on bitmap templates (1024 x 786 pixels). The X-Y fixations were then converted into distinct codes for the characters and background so that participants' fixations were mapped onto the objects of an image (the background, the pilot, the wizard, the detective, and the distractor objects - a tree and a bag - for Fig. 1). We used exact object regions for the analyses. This means that if a fixation did not occur exactly on the object, it was counted as a fixation on the background. Characters were coded depending on their event role for the inferential and descriptive analyses. For Fig. 1, for instance, and sentences (a1/a2) (Table 1) the pilot would be coded as 'patient', the wizard as 'stereotypical agent', and the detective as 'depicted agent'. Consecutive fixations within one object region (i.e., before a saccade to another region occurred) were added together and counted as one *inspection.* Contiguous fixations of less than 80 ms were pooled and incorporated into larger fixations; blinks and out-of-range fixations were added to previous fixations.

We rely on measuring eye movements to objects in a scene since previous studies on auditory sentence comprehension in visual environments have shown this measure to be closely linked to on-line comprehension processes (e.g., Sedivy et al., 1999; Tanenhaus et al., 1995). We report the percentage of inspections on the object regions, and inferential analyses of the number of inspections for individual time regions (Tables 3 and 5).

The inspection percentages for the individual time regions are based on exact computations of these regions for each individual trial. Word onsets had been marked for the first noun phrase, the verb, the adverb, and the second argument in each item speech file. Table 2 shows the extent and duration of the analysis regions for each condition. An inspection was counted as an inspection during a given region if it started within that time-region. To illustrate this: If an inspection of the depicted agent started after the onset of the verb and lasted into the adverb time region, it would be counted as an inspection of that agent during the verb, not, however, during the adverb region. If an inspection on the depicted agent started after the onset of the verb, ended before the end of the verb, and then a new inspection of the stereotypical agent started still before the end of the verb, then we record two inspections during the verb time region: one on the depicted agent, and one on the stereotypical agent.

For a time region, the percentage of inspections to an object in a condition was computed by dividing the number of inspections to that object in that condition by the total number of inspections to objects for that condition and multiply the result by 100. The inspection percentages thus sum to 100 within one condition for all regions, not however, for one object across the four conditions. This means that the relationship between the cell frequencies on which the percentages are based is preserved in a linear fashion when comparing cells in the table of inspection percentages within a row (condition). However, when comparing inspection percentages to the same entity across two conditions (i.e., between two rows in the table), such a linear relationship is not preserved. While it is true that the percentage measure is not a constant linear function for all the comparisons between entities, it is still informative of the relative inspections to one character over another for a given condition.

For the inferential analysis of inspection counts during a time region we used hierarchical log-

Table 2: *about here #*

linear models. These combine characteristics of a standard cross-tabulation chi-square test with those of ANOVA. Log-linear models neither rely upon parametric assumptions concerning the dependent variable (e.g., homogeneity of variance), nor require linear independence of factor levels, and are thus adequate for count variables (Howell, 2002). For the inferential analysis, inspection counts for a time region in the unique (Table 1, a1 and a2) and in the ambiguous identification conditions (Table 1, b1 and b2) were adjusted to factor combinations of target agent (depicted agent, stereotypical agent), target type (depicted, stereotypical), and either participants (N = 40) or items (N = 24). We report effects for the analysis with participants as $LR\chi^2(subj)$ and for the analysis including items as a factor as $LR\chi^2(item)$.

Results and Discussion

Tables 3 and 5 show the percentage of inspections to the characters (background, patient, stereotypical agent, depicted agent, distractors) during the four word regions (NP1, VERB, ADV, NP2), and Table 4 presents the results of the corresponding log-linear analyses. The relatively high percentage of inspections on the background presumably results from the fact that no scene was present during utterance presentation. Inspections to the regions that characters occupied on the previously-inspected scene would thus expectedly be less precise than for an immediately present scene.

We first present findings in the unique identification conditions to see whether people can use depicted events and stereotypical thematic role knowledge when they are uniquely identified as relevant by the utterance (unique identification conditions). Subsequently, we present findings when the verb identifies both target agents as relevant to see whether there is a priority of either depicted events or stereotypical thematic role knowledge (ambiguous identification conditions).

Unique identification conditions

Analyses of inspection counts in the NP1 and VERB regions showed that there were no significant effects for the analyses variables. Log-linear analyses for the ADV region revealed a significant interaction between target character (depicted agent, stereotypical agent) and target type (depicted, stereotypical target) in the participant analysis (see Table 4). Loglinear contrasts confirmed that there was a higher percentage of inspections to the stereotypical agent in the stereotypical (a2) than in the depicted (a1) target condition ($LR\chi^2(subj) = 18.71, df = 1, p < 0.001; LR\chi^2(item) = 20.53, df = 1, p < 0.0001$), and more inspections to the depicted agent in the depicted target (a1) than in the stereotypical (a2) target condition ($LR\chi^2(subj) = 19.11, df = 1, p < 0.001; LR\chi^2(item) = 17.69, df = 1, p < 0.0001$). A main effect of more inspections to the depicted agent than to the stereotypical agent was reliable with items as a factor (see Table 4).

Table 3: *about here #*

Table 4: *about here #*

For the NP2 region, there was a significant interaction between target character and target type in the absence of a significant main effect of target character (Tables 3 and 4). Contrasts confirmed that the interaction resulted from more inspections to the stereotypical agent in the stereotypical (a2) than in the depicted target condition (a1) ($LR\chi^2(subj) = 15.83, df = 1, p < 0.001; LR\chi^2(item) = 21.91, df = 1, p < 0.0001$), and a higher number of inspections to the depicted agent in the depicted target (a1) than in the stereotypical target condition (a2) ($LR\chi^2(subj) = 17.38, df = 1, p < 0.001; LR\chi^2(item) = 19.92, df = 1, p < 0.0001$).

Ambiguous identification conditions

For the NP1 and the VERB region there were no significant effects for the analyses variables. For the ADV region we found - as expected - no interaction since sentences were identical prior to the second noun phrase. Rather, we replicated the main effect of significantly more inspections to the depicted than to the stereotypical agent for both (b1) and (b2) that was found by Knoeferle and Crocker (2006) (see Tables 4 and 5). These looks occurred after people had heard 'The pilot (object/patient) spies-on ...' and before they heard the respective second noun phrase, which then disambiguated towards the depicted or stereotypical target type. Log-linear analyses showed that the main effect of a higher number of inspections on the depicted agent for the depicted target compared with the stereotypical target condition was significant.

Table 5: *about here #*

For the NP2 region, the interaction between target character (depicted, stereotypical) and target type (depicted target, stereotypical target) was reliable in the item analysis only. For a later NP2 region, however, (from the onset of the second noun phrase until 500 ms after its offset), the interaction between target character and target type was clearly reliable. For completeness sake, we present the descriptive data (Table 5) and the log-linear results for this later region in both the unique and ambiguous identification conditions (Table 4). For this later NP2 region, contrasts between a higher number of inspections to the depicted agent in the depicted compared with the stereotypical target condition were only reliable in the item analysis ($LR\chi^2(item) = 21.86, df = 1, p < 0.000$). Contrasts for more inspections to the stereotypical agent in the stereotypical versus the depicted target condition were also reliable in the item analysis ($LR\chi^2(item) = 27.98, df = 1, p < 0.000$). There was further a significant main effect of target character (Table 4) that resulted from a higher number of inspections to the depicted than stereotypical agent.

The key finding is that the analyses of the eye-movement data provide strong evidence for the rapid use of (non-stereotypical) depicted events, and for their greater relative importance over stereotypical knowledge of (non-depicted) events during online comprehension. Importantly, using the blank screen method, we replicate both the overall gaze patterns and the time-course of the findings by Knoeferle and Crocker (2006). The gaze pattern analyses thus support the view that the immediate visual presence of the entire scene is not a pre-requisite for the use and greater relative importance of depicted events.

In contrast with findings by Knoeferle and Crocker (2006), the difference between the interaction in the unique identification conditions and the main effect in the ambiguous identification conditions was not significant: The three-way interaction between target character (depicted agent, stereotypical agent), identification (unique, ambiguous), and target type (stereotypical, depicted) was not reliable. The failure to find this interaction may result from the overall weaker main effect and a consequently smaller difference between the main effect in the ambiguous condition and the interaction in the unique identification condition: Differences in inspections to the depicted versus stereotypical agent (10.0 percent for the depicted and 7.1 for the stereotypical target conditions) were only half as large as for findings by Knoeferle and Crocker (2006) (20.5 percent for the depicted and 18.4 for the stereotypical target conditions). The fact, however, that we replicate both a reliable interaction for the unique identification conditions and a significant main effect for the ambiguous conditions even for the blank screen paradigm provides strong support for the claim that people prefer to rely on depicted events over their stereotypical thematic role knowledge for incremental thematic interpretation, even when the scene is no longer present.

Experiment 2

Findings from Experiment 1 have shown that having to access non-stereotypical depicted events from working memory neither prevents their use, nor diminishes the preference to rely on them over stereotypical thematic role knowledge for incremental thematic interpretation. Examining the effects of event co-presence with the blank screen method may, however, have disadvantaged the stereotypical agent as much as, or potentially even more than, the agent of the depicted event: In real-world events, the action of an agent upon another event participant is often time-limited and only briefly visible. One consequence is that action events do not necessarily occur simultaneously with a related utterance but sometimes only partially overlap, precede or follow it. The participants of an event, in contrast, would often be present before and after performing an action. Furthermore, while it appears felicitous to exploit a previously completed action that is no longer co-present for comprehension, anticipating a stereotypical event that has not yet been depicted may have been infelicitous in Experiment 1 since the screen did not contain a character that could, in principle, perform a future action.

Experiment 2 addresses these considerations: While non-stereotypical events are not depicted during utterance comprehension, a stereotypical agent is co-present and available for potential future events. To achieve this, Experiment 2 briefly depicted and removed the actions while the characters remained as listeners heard the utterance.

If people rapidly rely on either stereotypical thematic role knowledge or on depicted events when they are uniquely identified as relevant, then we would expect to replicate the gaze patterns observed in the unique identification conditions of Experiment 1 and Knoeferle and Crocker (2006, Exp. 2): more inspections to the stereotypical agent in the stereotypical than in the depicted target condition, and more inspections to the depicted agent in the depicted than in the stereotypical target condition, shortly after the verb.

Further, if the relative priority of depicted events prevails even when the events are only briefly presented and absent during utterance comprehension while a competing stereotypical agent remains, this would suggest that the visual absence of depicted events and their potentially completed aspect has no bearing on their relative priority. In this case we would expect to find the same gaze pattern and time-course of effects as in the ambiguous conditions of Experiment 1 and Knoeferle and Crocker (2006, Exp. 2). In contrast, a failure to replicate the depicted-events priority in the ambiguous conditions would suggest that the absence - and possibly completed aspect - of depicted events leads to a decay of the event representations in working memory. In contrast, the stereotypical agent is both co-present and supported through expectations generated by the verb.

Method

Participants

Forty further participants from the same population as in Experiment 1 were paid five euros for taking part in the experiment.

Materials, Design, Procedure and Analysis

Materials and design were identical, and presentation, analysis, and instructions were similar to Experiment 1: The only difference in procedure between Experiments 1 and 2 concerned image presentation. Rather than blanking the entire clipart scene as in Experiment 1, we presented the depicted actions in a sequence of four frames in Experiment 2. For the first three frames (Figs 2a-c) no speech input was presented concurrently. The initial scene showed the characters, not however, the instruments that depicted the actions (Fig. 2a). Participants inspected the scene in Fig. 2a for 1500 ms. In the second and third frames, first, one action (e.g., wizard, spying with a telescope, Fig. 2b) and 1500 ms later the action of the other agent (detective, serving food on a plate, Fig. 2c) was presented. After another 1500 ms, both actions and any instruments disappeared together, the characters remained, and after a further 1500 ms the utterance was presented while people inspected the scene in Fig. 2d. To balance directionality effects, the action appeared first for the agent left of the patient, and subsequently for the agent to the patient's right for one half of the items; for the other half, the action appeared first for the agent to the right, and then for the agent to the left of the patient. To examine whether the order of action appearance influenced gaze patterns in the final scene, the directionality was included as a factor in the analyses (left-right vs. right-left).

Figure 2: *about here #*

A further difference between Experiments 1 and 2 concerned the analysis templates that permitted mapping of the X-Y fixation coordinates onto scene objects. In Experiment 1, the templates were based on Fig. 1a since participants inspected this image prior to utterance presentation. For Experiment 2, however, templates - based on the final scene in Fig. 2d - did not depict actions but only contained the characters as regions (background, patient, stereotypical agent, depicted agent, distractors).

Since we used a rather different, and more dynamic presentation prior to utterance comprehension in Experiment 2, we provide an overview of the time course of gaze patterns to the target agents in the unique (Fig. 3) and ambiguous identification conditions (Fig. 4) in addition to the more detailed word region analyses. For the time course presentation of the eye-movement data we computed the proportion of inspections per condition that fell within a given time slot for each object (see Knoeferle & Crocker, 2006). The word onsets marked on the graphs (Figs. 3 and 4) represent the average of word onsets for the item trials.

Finally, there was a minor change in the instructions (replacing 'previously viewed scenes' with 'previously viewed sequence of scenes') to accommodate the fact that the events were no longer depicted on a previously viewed scene but rather over a sequence of scenes.

The filler images and sentences were the same as for Experiment 1. Presentation was, however, changed to resemble that of the experimental images. Thus, for each filler trial, participants saw a sequence of four related scenes on which objects and characters were briefly presented, moved, or were removed.

Results and Discussion

Figs 3 and 4 show the time course of gaze pattern to the target characters (depicted agent, stereotypical agent), and Tables 6 and 8 the percentage of inspections to the target characters (background, patient, stereotypical agent, depicted agent, distractors) during the analysis regions (NP1, VERB, ADV, NP2). We present the analyses of gaze patterns first for the unique (Fig. 3 and Table 6) and subsequently for the ambiguous identification conditions (Fig. 4 and Table 8).

Unique identification conditions

The time course of inspections during comprehension suggests that during the first noun phrase, the gaze pattern was somewhat noisy with more inspections to the stereotypical agent in the depicted than in the stereotypical target condition and vice versa for the depicted agent. Shortly after people encountered the verb, however, the time course data provide clear support for the rapid utterance-mediated use of either depicted events or stereotypical thematic role knowledge. People inspected the depicted agent more frequently in the depicted than stereotypical target condition, and the stereotypical agent more often in the stereotypical than depicted target condition (Fig. 3).

Inferential analyses presented in Table 7 confirmed these findings. For the NP1 region there was an interaction of target character (depicted agent, stereotypical agent) and target type (depicted, stereotypical) that was marginal in the analysis with participants and significant in the item analysis. We discuss the effects on the first noun phrase at the end of the 'Results' section. For the VERB region, there was an interaction of target agent with target type, in the absence of a main effect of target character.

Figure 3: *about here #*

During the ADV region, there was an interaction between target character (depicted agent, stereotypical agent) and target type (depicted, stereotypical) that was marginal in the analysis with participants and significant in the item analysis while the main effect of target character was not reliable.

Table 6: *about here #*

Table 7: *about here #*

While this interaction was not that strong on the adverb region, when looking at a more extended region ('VerbAdv') that started at verb onset and lasted until the offset of the adverb for each trial, the interaction between target character and target type was clearly reliable (Tables 6 and 7). Contrasts for this extended VerbAdv region confirmed that the interaction resulted from a higher percentage of inspections to the stereotypical agent in the stereotypical (a2) than the depicted target condition (a1) ($LR\chi^2(subj) = 4.60, df = 1, p < 0.05; LR\chi^2(item) = 3.94, df = 1, p = 0.05$), and an increased number of inspections to the depicted agent for the depicted (a2) than the stereotypical target condition (a1) $LR\chi^2(subj) = 14.14, df = 1, p < 0.001; LR\chi^2(item) = 13.56, df = 1, p < 0.001$) (Table 6). Recall that we included directionality of action appearance (left-right vs. right-left) in the analyses to see whether they affected gaze patterns. Interactions between directionality (left-right, right-left), target character (stereotypical agent, depicted agent), and target type (stereotypical, depicted) as well as between target character and directionality were not significant.

Analyses of inspection counts for the NP2 region found a significant interaction between target character and target type. Contrasts again confirmed that there was a higher number of inspections to the stereotypical agent in the stereotypical than in the depicted target condition ($LR\chi^2(subj) = 26.14, df = 1, p < 0.0001; LR\chi^2(item) = 24.80, df = 1, p < 0.0001$), and that there were more inspections to the depicted agent in the depicted than in the stereotypical target condition ($LR\chi^2(subj) = 21.55, df = 1, p < 0.0001; LR\chi^2(item) = 18.65, df = 1, p < 0.0001$). Interactions between directionality (left-right, right-left), target character (stereotypical agent, depicted agent), and target type (stereotypical, depicted) as well as between target character and directionality were not significant.

Ambiguous identification conditions

For the ambiguous conditions, the time curves (Fig. 4) suggest that - unlike in Experiment 1 - people inspected the agent depicted as performing the action no more frequently than the stereotypical agent shortly after the verb.

Inferential log-linear analyses are presented in Table 7. For the NP1 region, log-linear analyses confirmed that the interaction between target character (depicted agent, stereotypical agent) and target type (depicted, stereotypical) was reliable. For the VERB region, there was neither a reliable main effect of target character nor a reliable interaction of target character and target type.

Figure 4: *about here #*

For the ADV region there was - as expected - no interaction of target character and target type since sentences were identical prior to the second noun phrase. However, unlike in Experiment 1 and Knoeferle and Crocker (2006, Exp. 2), we also find no main effect of target agent (Table 8).

For the more extended VerbAdv region, analyses confirmed there was no interaction of target character and target type, but a marginal main effect of target character that resulted from slightly more inspection of the stereotypical rather than depicted agent. Crucially, there was no reliable interaction between directionality (left-right, right-left), target character (stereotypical agent, depicted agent), and target type (stereotypical, depicted) as well as between target character and directionality.

For the NP2 region, there was a reliable interaction between target character and target type in the absence of a significant main effect of target character. Contrasts for the comparison be-

Table 8: *about here #*

tween inspections to the stereotypical agent in the stereotypical compared with the depicted target condition ($LR\chi^2(subj) = 9.31, df = 1, p < 0.01; LR\chi^2(item) = 8.22, df = 1, p < 0.01$) and for inspections to the depicted agent in the depicted versus the stereotypical target condition were reliable ($LR\chi^2(subj) = 25.18, df = 1, p < 0.0001; LR\chi^2(item) = 27.60, df = 1, p < 0.0001$). Interactions between directionality (left-right, right-left), target character (stereotypical agent, depicted agent), and target type (stereotypical, depicted) as well as between target character and directionality were not reliable.

The key findings of Experiment 2 are that people, just as in Experiment 1, still rapidly *use* depicted events; however, unlike in Experiment 1, they no longer *prefer* to rely on them: For the unique identification conditions, analyses of gaze patterns during the ADV and VerbAdv regions replicated the finding that people rapidly rely on either stereotypical thematic role knowledge or on depicted events in anticipating a relevant role filler (Knoeferle & Crocker, 2006, Exp. 2, and Experiment 1 of the present paper) even when the actions were only briefly depicted and then removed prior to utterance presentation while the characters remained.

In the ambiguous conditions, when the verb ('spies-on') identified both a stereotypical (the detective) and a non-stereotypical agent of a depicted spying action (the wizard) as relevant, analyses of gaze patterns confirmed that there was no longer a preference to anticipate the agent that is depicted as performing the spying action (the wizard). Rather, we saw a tendency towards anticipating the stereotypical agent more than the agent depicted as performing the action during the VerbAdv region.

Importantly, the three-way interaction between target character (depicted agent, stereotypical agent), identification (unique, ambiguous), and target type (stereotypical, depicted) on the VerbAdv region was significant in the participant analysis, thus confirming that the difference between gaze patterns in the unique and ambiguous conditions was reliable.

Unlike Experiment 1 and Knoeferle and Crocker (2006, Exp. 2), we found a reliable interaction between target agent and target type on the first noun phrase despite the fact that presentation was identical across conditions up to this region. The interaction resulted from more inspections to the stereotypical agent in the depicted than in the stereotypical target condition, and more gazes to the depicted agent in the stereotypical than in the depicted target condition (thus, people always inspected the inappropriate character for a condition). There were further more inspections to the target characters in Experiment 2 than Experiment 1. We think that these differences in gaze patterns may be the result of the differences in presentation between Experiments 1 and 2: Shortly before hearing the first noun phrase, people had seen an action appear one after another, and then disappear together. This may have prompted people to inspect both agents more often than in Experiment 1. Further, it may have increased the noise in the eye movements at the onset of the utterance. Log-linear analyses that included all of the scene objects in the analysis provided further support for this view by showing that the interaction on NP1 between target type and target character (background, patient, depicted agent, stereotypical agent, distractors) was not reliable in this more complete model.

The findings for Experiment 2 suggest that listeners can still rapidly exploit depicted event sequence information in the unique conditions, despite the fact that event information is not visually co-present during utterance comprehension. The relative priority of that event information, however, was eliminated in the ambiguous identification conditions.

Experiment 3

There are two possible explanations for the diminished priority of event information in Experiment 2. A decay account suggests that while prior event information can still be exploited (unique conditions), decay weakens its salience such that when a plausible competitor is co-present, there no longer is a priority of depicted events (ambiguous conditions). An aspectual account of the findings from Experiment 2, in contrast, suggests that listeners, when faced with a choice between relying on an event that has been depicted as completed versus hypothesizing a future event, if anything, have a slight preference to anticipate the plausible role filler of a future event. This bias may further have been facilitated by the future tense of the utterance in our materials.

The main goal of Experiment 3 was to see whether the rapid use of scene events would be re-established when depicting them in sequences similar to those of Experiment 2, but with plausible rather than non-stereotypical role fillers. There were thus always two objects in the scene that were equally plausible targets for an action indicated by the verb. One of these objects was acted upon prior to utterance comprehension while the other object did not undergo an action. At the verb, people had a choice between anticipating the plausible object that had already been acted upon or an equally plausible object that had not yet undergone an action.

Keeping plausibility of the role fillers that are associated with a depicted and a non-depicted event constant, will enable us to distinguish between a decay versus aspectual account. Prior depicted events are now the only factor differentiating the two candidate role fillers. The decay account predicts that since both role fillers are equi-plausible, listeners will prefer the role filler associated with the decayed depicted event. If, however, people generally disfavour event information that has been perceived as completed, then we would expect no such bias towards the use of the completed event, thus supporting the aspectual account.

In addition, we included a tense manipulation: utterances either referred to a future or a past event. In Experiments 1 and 2, which had kept the sentence stimuli from Knoeferle and Crocker (2006), sentences always referred to an event in the immediate future (e.g., 'spies-on soon', Table 1). This was a felicitous manner of identifying an event as relevant in static scenes. For a more dynamic presentation of events prior to comprehension, however, people may perceive the actions as completed. Referring to the past depicted events in a felicitous way may make them more accessible. Equally, future tense might bias towards the expectation of a future event, leading to anticipation of an object that has not yet undergone an action. In previous studies with co-present static scenes, gaze patterns have indeed provided some evidence that people rely on tense cues for anticipating relevant stereotypical role fillers (e.g., Kamide et al., 2003) as well as for accessing depicted events (Knoeferle et al., 2005; Knoeferle & Crocker, 2006), although their influence is typically weak.

Finally, the fact that Experiment 3 used completely new stimuli further allowed us to examine whether the findings for non-canonical object-verb-subject sentences and non-stereotypical agent-action-patient events by Knoeferle and Crocker (2006) generalize to other sentence types such as canonical subject-verb-object (SVO) sentences and more natural scenarios such as plausible agent-action-theme events.

Method

Participants

Further fifty-six participants from the same population as in Experiment 1 received five euros for participating in the experiment.

Materials, Design, Procedure, and Analysis

There were sixteen items, and two conditions (Fig. 5 and Table 9 (a) and (b)). As a result of counter-balancing, however, an item comprised two sequences of three clipart scenes, and four sentences (Fig. 5 and Table 9). We first describe the stimuli, presentation, and design, and subsequently the counter-balancing.

An example Ersatz scene showed an agent (a waiter), several chandeliers, a set of crystal glasses, and distractor objects (e.g., Fig. 5a). Two objects in the scene (e.g., the chandeliers and the crystal glasses) were plausible themes for the same event (e.g., waiter-polishing). The verb in the utterance (e.g., *polieren* 'polish') always identified both of these objects as relevant (Table 9 (a) and (b)).

Figure 5: *about here #*

Only one of these two target objects, however, was acted upon prior to utterance presentation (e.g., the chandeliers, Table 9 (b)). Participants inspected the scene in Fig. 5a for 1500 ms. On the next frame, the waiter was depicted as having moved to the chandeliers, and polishing them (Fig. 5b). After 1500 ms a third image was presented on which the action (and any instruments) had been removed, and the waiter had moved to a new and inactive position (Fig. 5c). The third image remained on the screen and 1500 ms after display onset for Fig. 5c the utterance was presented. Plausibility of the events together with the presentation and the ambiguity at the verb were crucial features since they allowed us to see whether - for plausible events that were depicted as a sequence of actions - people would rapidly anticipate the object that had previously been acted upon even when the verb ambiguously identified two objects as relevant themes.

Table 9: *about here #*

After the action sequence, one of the objects (the chandeliers) had thus become the target of a past event. The other plausible object (the crystal glasses) had not yet been acted upon and could be the target of a future polishing event. While the verb identified both of these target objects as relevant, the utterances differed in whether they identified the object of a past or a future event as relevant (future (a) vs. past (b) tense condition, Table 9). For the future tense condition (a), the verb was in the present tense and the post-verbal adverb expressed a future bias, and in the past tense condition, verb and adverb were in the past. While it would have been desirable to use a future tense for the verb in the future condition (a), this was not possible for the present design since it - as outlined above - crucially relied on the referential ambiguity at the lexical verb prior to the second argument. Using a verb in the future tense, however, would have required a sentence-final verb, and an auxiliary (*wird*, 'will') in the position following the subject. The temporal adverbs were similar to those in Knoeferle et al. (2005): We used *baldigst* ('as soon as possible'), *sogleich* ('instantly'), *nachher* ('afterwards'), and *demnächst* ('soon') in the future tense condition, and *vorhin* ('a little while ago'), *soeben* ('just now'), *unlängst* ('recently'), and *kürzlich* ('recently') in the past tense condition. In the past tense condition, the second argument always referred to the object that had been acted upon (the chandeliers, Table 9 (a)) while in the future tense condition, the second noun phrase referred to the other plausible object that had not yet undergone an action (the crystal glasses, Table 9 (a)).

To balance for differences in visual salience and in the plausibility of the objects (the chandeliers versus the crystal glasses) as target of the action expressed by the verb ('polish'), each target object (the chandeliers, the crystal glasses) was once the target of the depicted polishing event (e.g., the chandeliers, Table 5, (b) and (a')) and once the target of the future polishing event. Instructions and procedure were identical to Experiment 2. The corresponding words in the four sentences of an item were matched for number of spoken syllables and frequency of lemmas (Baayen et al., 1995). Each participant saw an individually randomized version of one of four experimental lists. An experimental lists contained one condition of an item, an equal number of items per condition, and the same number of trials for each condition. Fillers were identical to Experiment 2. Consecutive item trials were separated from one another by at least one filler trial.

If recently depicted event information decays but can still be exploited when plausibility does not bias towards an alternative agent, then we would expect to once more find a great importance of recently depicted events in Experiment 3. People should anticipate the object that had been acted upon (the chandeliers, Fig. 5a-c) more often than the alternative plausible object. Alternatively, if the aspectual account is correct, and it was the completed aspect of the events that eliminated the priority of depicted events, then people should not anticipate the object of the completed event more often than the object of a future event that has not yet been depicted.

Finally, if the past versus future bias can make a past and future event respectively more accessible, then we would expect the following gaze patterns: People should anticipate the object that had been acted upon (the chandeliers in Fig. 5a-c) more often for the past (b) than future (a) tense sentences shortly after the tense information has become available in the utterance and prior to the second noun. Conversely, we would expect more anticipatory inspections to the other plausible object that had not yet been acted upon (the crystal glasses, Fig. 5a-c) in the future (a) than past (b) tense condition. Such gaze patterns shortly after the verb and adverb and prior to the second argument would confirm that people can rely on either depicted events to anticipate the target of the depicted event or on their knowledge of plausible non-depicted events in anticipating the target of a future event. Such a finding would further provide support for the role of tense and aspect.

Analysis

The objects on the final scene (e.g., Fig. 5c) were coded depending on their role for the inferential and descriptive analyses. For Fig. 5c, for instance, and sentence (a) (Table 9) the waiter was coded as ‘agent’, the crystal glasses as ‘future target’, the chandeliers as ‘past target’. We report the percentage of inspections and inferential analyses of the number of inspections for individual time regions (Table 11). An inspection was counted for a time-region if it started within that time-region. Word onsets had been marked for the first noun phrase, the verb/auxiliary, the adverb, and the second argument in each item speech file. Table 10 shows the extent and duration of the analysis regions for each condition.

Table 10: *about here #*

Results and Discussion

Table 11 shows the percentage of inspections to the characters (background, agent, past target, future target, distractors) for the four analysis regions (NP1, VERB, ADV, NP2). We first briefly describe the gaze pattern and then follow up these descriptions with the inferential analyses of inspections to the target objects (see Table 12).

During the first noun phrase, people looked mostly at the agent and approximately equally often at the two target objects (the chandeliers, the crystal glasses). Once the verb is encountered, looks to the agent start to decrease, and people inspect the object that had previously been acted upon most frequently (Table 11). During the adverb, there are some effects of the tense manipulation, but differences are relatively small: People look more often at the future target in the future than in the past tense condition, and they inspect the past target more often in the past than in the future tense condition. This gaze pattern continues during the second noun phrase.

Log-linear analyses showed that there was no significant interaction between tense condition (future tense, past tense) and inspection of the target objects (past target, future target) for the NP1 region. For the VERB region, there was a main effect of target object in the absence of a reliable interaction between tense condition and target object. The main effect resulted from a higher number of inspections to the past target than to the future target in both tense conditions (Table 12). During the ADV region, the interaction between tense condition and target object was not reliable. There was further a main effect of target object (Table 12).

Table 11: *about here #*

Table 12: *about here #*

For the NP2 region, analyses confirmed a significant interaction between target object and tense condition only in the analysis with items as a factor. However, for a slightly later time region (from 400 ms after NP2 onset to 100 ms after the offset of NP2 for each trial), log-linear analyses confirmed the interaction between target object and tense condition. Contrasts confirmed that the interaction resulted from a higher number of inspections to the future target in the future than in the past tense condition. Contrasts for inspections to the past target in the past versus future tense condition were not reliable.

The gaze pattern analyses crucially showed that people had a preference to inspect the object that had been acted upon (the past target, Table 11) more often during the verb than the alternative plausible object (the future target). This finding provides evidence for the decay account, and against an aspect account of the findings in Experiment 2, clearly revealing the importance of (previously) depicted events when role-fillers are equi-plausible but events are presented as completed (as in Experiment 2). Further, the effects of tense were not reliable. The preferred anticipation of the acted-upon target, even in the future tense condition of Experiment 3, provides further evidence against an "aspectual" account for the findings in Experiment 2, in which the future tense potentially made previously depicted events less accessible. Analyses during the second noun phrase finally confirmed that people correctly interpret the utterance and mostly inspect the mentioned object.

While the findings of Experiment 3 provide further support for the priority of plausible, dynamically depicted events, the extent to which our findings in turn generalize to more realistic or real-world video scenes remains to be examined (see Almeida, Nardo, & Grunau, 2006).

General Discussion

We have reported findings from three experiments that examined the relative priority and use of depicted events when the events in an Ersatz scene were not visually co-present, and only their representations in working memory could be accessed during utterance comprehension. Taken together, findings from the three studies importantly reinforce previous claims concerning the strong relative importance of depicted events for comprehension processes, and the rapid time-course with which such scene information is applied. Importantly, however, our findings provided evidence that the priority of depicted events can be modulated by decay when the events are no longer co-present. The findings from these three experiments motivate a revision of the original Coordinated Interplay Account which incorporates an explicit working memory. We first discuss the experimental findings, and subsequently consider their implications for the Coordinated Interplay Account.

Experiment 1 found that the visual co-presence of the entire scene is not a pre-requisite for the use and priority of depicted events that was observed by Knoeferle and Crocker (2006). Gaze patterns in the unique condition confirmed that people were able to use either verb-based stereotypical thematic knowledge or non-stereotypical depicted events when the verb uniquely identified either a stereotypical agent or the agent of a previously-depicted action respectively as relevant for thematic interpretation. Furthermore, when the verb identified two different agents as relevant, we replicated the finding that people prefer to rely on the agent that had been depicted as performing the action over the stereotypical agent for thematic interpretation. The eye movements that supported these claims occured - just as in Knoeferle and Crocker (2006) - shortly after the verb and prior to the second noun phrase. However, unlike in their experiments they occurred to regions in the blank screen that had previously been occupied by events rather than to co-present depicted events. The findings entail that we extend the Coordinated Interplay Account with an explicit notion of working memory that enables the comprehension system to rapidly exploit information about depicted events that are no longer co-present.

The data further confirmed that the gaze patterns and time course observed by Knoeferle and Crocker (2006) generalize to the blank-screen paradigm (see Altmann, 2004; Spivey & Geng, 2001). In this respect, our findings extend insights into the access of object properties by Spivey et al. (2004) and Altmann (2004) to more complex scenes containing events. Clearly, the flexible manner in which people use recently presented scene information confirms that the “visual world” paradigm is robust even when scenes are not co-present.

Findings from Experiment 2 provide more detailed insights into how action presentation affects the priority of non-stereotypical depicted events over stereotypical thematic role knowledge. Actions were briefly depicted and removed, approximating more natural sequences of past, and possibly completed, events than the blank-screen method. This presentation further disadvantaged only the depicted events since the characters remained visible during comprehension and were available for performing future actions. Gaze patterns confirmed that people were still able to use either verb-based stereotypical thematic knowledge or non-stereotypical depicted events when the verb uniquely identified these information types as relevant. However, when the verb identified two different agents as relevant, we no longer found a preferential anticipation of the agent that had previously been depicted as performing an action. Rather, there was a tendency to anticipate the stereotypical agent more often than the agent of the (non-stereotypical) depicted event, but this effect was not reliable.

Together, the analyses of gaze patterns from Experiments 1 and 2 suggest that the visual co-presence of an event is not required to explain the findings by Knoeferle and Crocker (2006). If visual co-presence were crucial, we should not have replicated the findings in the blank-screen method (Experiment 1), and we would also not have expected that people still rely on depicted events in the unique condition of Experiment 2. Rather, the findings point to an account which highlights the role of working memory: Specifically, that prior scene information can inform subsequent comprehension, but also that objects and events that are no longer co-present during comprehension - or possibly perceived as completed - experience some decay.

Experiment 3 provided additional support for this account. Using the same action-sequence presentation as Experiment 2, but with newly-created plausible agent-action-theme relationships instead of non-stereotypical events, a rapid use of event information was observed. When the verb identified both the co-present plausible target of a previously-depicted event as relevant or an alternative co-present plausible object that had not yet been acted upon, people fixated the object that had previously been acted upon more often than the other plausible object in the scene. This finding argues against an aspectual explanation for the diminished importance of (completed) events in Experiment 2, since the target of the depicted event in Experiment 3 is favoured despite a similar presentation of the events as completed. Rather, findings support the decay account in which completed events are still exploited but have somewhat weaker influence due to decay. The decayed event in Experiment 2 cannot overcome the competing bias towards the stereotypical agent. In Experiment 3, however, when plausibility is held constant, the decayed event is sufficient to dominate interpretation processes. We further investigated the time-course with which tense cues might bias people towards either accessing a past (depicted) event or interpreting the verb as indicating a future (non-depicted) event. While gaze pattern provided some evidence for the influence of tense cues on inspections to either a past or future event target, effects were not reliable. The fact that these effects were relatively small is consistent with previous findings by Knoeferle and Crocker (2006, Exp. 1).

The findings from Experiments 2 and 3 strongly suggest that the relative importance which the comprehension system accords to depicted events for incremental thematic interpretation depends on their accessibility (from the co-present scene versus working memory) and the extent to which event representations in working memory have decayed. Implications for, and a revision of, the Coordinated Interplay Account are discussed below.

The Revised Coordinated Interplay Account

Findings from these three experiments necessitate a refinement of the original Coordinated Interplay Account. Construed narrowly, the original account treated the verb-mediated visual inspection of a co-present depicted event as essential for the priority of that event. The findings from the present three experiments, however, clearly showed that people rapidly use depicted events for thematic role assignment even when they are no longer visually co-present.

Accounting for both our findings (Experiments 1 to 3) and those of others (Richardson & Spivey, 2000; Spivey & Geng, 2001; Altmann, 2004), thus requires an explicit notion of working memory. It is this memory which allows information from depicted events that are no longer co-present to be accessed during comprehension. One important question is how to implement the working memory mechanism: Gaze patterns in the blank screen revealed that the absence of the events does not disrupt the interplay between comprehension, attention, and the temporally coordinated use of scene information, supporting the view that people rely on pointers from an object's or event's working memory representation to the location of the object/event in the scene in accessing associated object/event properties. These pointers are still accessed in a closely time-locked manner by visual attention mechanisms, a finding that is consistent with the original Coordinated Interplay Account. Attention in the scene thus retains its central role in the model. However, while the original account assumed that the search for potential referents must occur through utterance-mediated attention in a co-present scene, the revised account offers a second path: In addition to the scene, comprehension can also consider working memory representations of scene objects and events in the search for an appropriate referent.

Further to the evidence for the rapid exploitation of scene event representations, an important implication of our findings is that the activation of such working memory representations may decay, presumably leading to a decrease in their accessibility and importance for comprehension. The descriptively smaller effect for the priority of depicted events when the screen was blank (Experiment

1) compared with the original study in which a scene was co-present (Knoeferle & Crocker, 2006) is suggestive of such decay. The proposal receives further support from Experiment 2, when simply keeping a stereotypical agent in the scene while the depicted events were absent, eliminated the priority of depicted events (though the decayed event representations were still used in the unique condition). Findings from Experiment 3, finally, make clear that once the target of the completed depicted event is a plausible theme of the verb, triggering its anticipation, then the decayed depicted action, combined with the plausibility of the target, results in a preference to rely on the event.

Fig. 6 presents a formalized outline of the revised Coordinated Interplay Account, detailing the time course of processes and their informational dependencies. While outlining key processes at word i in a quasi-serial order, we embrace the possibility that they may partially overlap and occur in parallel, modulo informational dependencies. There are three central steps (*i*, *i’*, and *i”*) for the processing of word$_i$. The state of working memory in our model is indicated in the ‘WM’ box: It includes scene- and utterance-based representations for each step of processing word$_i$. The contents of working memory and the co-present scene objects/events are updated at each step using the prime (’) notation. ‘Sentence Interpretation’ (step i) indicates that word$_i$ is integrated into the existing interpretation int$_{i''-1}$ (yielding int$_i$), and that linguistic expectations (ant$_i$) are derived in the process based on the new interpretation, linguistic/long-term knowledge, and previous expectations ant$_{i''-1}$. Working memory for step i includes the interpretation (int$_i$), linguistic expectations (ant$_i$), and the previously inspected scene (scene $_{i''-1}$). ‘Utterance Mediated Attention’ (step *i’*) details the search process for a referent both in the scene itself and in working memory representations of the scene. Newly attended scene information is added into the prior working memory representations of the scene (scene$_{i''-1}$), yielding scene$_{i'}$. Representations of objects and events that are no longer in the scene then decay. ‘Scene Integration’ (step *i”*) outlines the reconciliation of the interpretation int$_{i'}$ with scene$_{i'}$ from working memory$_{i'}$. The reconciliation step consists of a coindexing process (to establish reference) and any revisions necessary based on scene events. In addition, the linguistic expectations ant$_{i'}$ are reconciled with scene$_{i'}$. Working memory for step *i”*contains the revised interpretation (int$_{i''}$), expectations (ant$_{i''}$), and scene-based representations (scene$_{i''}$). The contents of working memory$_{i''}$ are then available to inform the new interpretation (int$_{i+1}$) and linguistic expectations (ant$_{i+1}$) when word$_{i+1}$ is heard.

Figure 6: *about here #*

To illustrate both the added working memory mechanism and the notion of working memory decay, we outline how the revised Coordinated Interplay Model accounts for findings in the ambiguous identification conditions for all three experiments. Particular attention will be given to the support that the two competing targets (the agent of a non-stereotypical, depicted event versus the stereotypical agent of a non-depicted event in Experiments 1 and 2; the two plausible theme role fillers in Experiment 3) receive at the verb according to the notion of decay, linguistic expectations, and scene depiction.

Re-consider the example sentence *Den Piloten bespitzelt der Detektiv* (‘The pilot$_{acc/obj}$ spies-on the detective$_{nom/subj}$’) for Experiment 1 (see Fig 6). The state of working memory at step *i”-1*, after processing ‘pilot’ (word$_{i-1}$) and just before the verb is heard, is as follows: The interpretation (int$_{i''-1}$) contains a pilot that is the patient of an as-yet-unknown action. Scene-based representations (scene$_{i''-1}$) possibly include representations of events that were proximal to the pilot: wizard-spying-on-pilot, and detective-serving-food-to-pilot. When the verb (word$_i$, ‘spies-on’) is heard, its meaning is accessed, and integrated into the previous interpretation (int$_{i''-1}$). The result of this integration process is a new interpretation (int$_i$), containing a spying action of which the pilot is the patient.

At the verb, linguistic expectations alone (ant$_i$) - derived from the new interpretation (int$_i$) and long-term knowledge - support a stereotypical agent of the spying action (e.g., the detective). Based on the new interpretation, and the current referring expression (‘spies-on’), the scene itself and scene-based representations in working memory are searched (step *i’*). Both the depicted events and the agents experience decay since the entire screen was blanked in Experiment 1. The representation of ‘spies-on’ matches a scene-derived representation for a spying action in working memory$_i$ (scene$_{i''-1}$). The representation of the depicted action further encodes the action’s location in the previous scene through a pointer to that location. Visual attention is directed to that location, presumably grounding referential processing in the scene. Reference is established, and coindexing creates a link between the meaning of the verbal referring expression (‘spies-on’), the referent’s previous location, and associated information about the referent. In inspecting the prior location of the referent, some attention presumably also goes to proximal scene locations, activating the memory representations for objects/events that previously occupied these locations. Information from working memory about the referent and previous proximal objects/events (scene$_{i'}$), is then available at step *i”*for reconciliation with the previous expectations (ant$_{i'}$) and for revising the interpretation (int$_{i'}$), yielding ant$_{i''}$ and int$_{i''}$ respectively. Accounting for the findings of Experiment 1 is thus achieved with the original Coordinated Interplay Mechanism and the working memory extension.

In Experiment 2, unlike Experiment 1, the agents are crucially co-present in the scene. The verb refers to a depicted action that is no longer co-present in the scene, but working memory still contains an appropriate, decayed, representation of a previously depicted action. The representation of the depicted action encodes the action's location in the previous scene through a pointer to that location. Visual attention is directed to that location, retrieving associated information about the action referent. Reference is established, and coindexing creates a link between the meaning of the verbal referring expression ('spies-on') and the referent's previous location. In inspecting the location of the referent, people notice the proximal, co-present, agent (the wizard). The agent of the depicted, non-stereotypical action thus receives support through being associated with a past action, however, the action representation has experienced some decay. The alternative, stereotypical co-present agent (the detective) has not been depicted as performing a spying action. Verb-based linguistic expectations of plausible events, however, support the detective as a stereotypical spying agent. In addition, representations of the stereotypical agent experience no decay since the agents are co-present in the scene. The decay of the depicted event versus the continued co-presence (and thus activation) of the stereotypical agent together with the support that the stereotypical agent receives from linguistic expectations, accounts for overriding the priority of depicted events in the model.

Consider a further example from Experiment 3: *The waiter*$_{subj/ag}$ *polishes soon the chandelier*$_{obj/pat}$. The visual context shows a waiter, a chandelier, and crystal glasses. The state of working memory after the first noun phrase (word$_{i-1}$) contains an interpretation (int$_{i''-1}$) with the waiter as the agent, and scene information (scene$_{i''-1}$) that presumably includes a waiter-polishing-chandelier event, and possibly the crystal glasses. The verb (word$_{i}$) is integrated with the previous interpretation (int$_{i''-1}$), yielding an interpretation (int$_{i}$) that contains a waiter involved in a polishing event. Linguistic expectations (ant$_{i}$) based on long-term knowledge and the new interpretation (int$_{i}$) support both the chandelier and the crystal glasses as plausible targets, and both are co-present during comprehension. People search the scene based on verb meaning ('polish'), but find no action referent. Working memory, however, contains a representation of a polishing event (step *i'*). Based on the pointer of that representation, the previous location of the completed polishing event is inspected, and reference between the meaning of the referring expression, the (decayed) working memory representation of the action, and its location in the scene is established. In inspecting the previous location of the event, people inevitably notice the object that had recently been acted upon (the chandelier). Having found a potential role filler associated with the action representation, the expectations (ant$_{i'}$), and the interpretation (int$_{i'}$), are reconciled with that scene information (scene$_{i'}$), and now include a waiter as the agent of a polishing action, and the chandelier as the most likely patient (step *i”*). At the next word, information from int $_{i''}$ as well as from ant$_{i''}$ is available for informing the new interpretation (int$_{i+1}$) and expectations (ant$_{i+1}$). With its working memory extension and the notion of decay in working memory representations, the model accounts for both the gaze pattern and the time course of the present findings.

The Relative Importance of the Scene

While accounting for the above findings through the coordinated interplay mechanism and decay in working memory, several other factors that have not yet been explicitly included into the Coordinated Interplay Account may modulate the use and relative importance of scene objects and events.

Among these are, for instance: the extent of referential success (see Introduction and Knoeferle and Crocker (2006)); locational or temporal cues in the utterance that clarify the (ir)relevance of the immediate scene; a shared versus different frame of reference between speaker and listener; the extent to which people have already inspected an object/event (involving both factors such as scene complexity and time), and the availability of social interaction (e.g., gestures) for explicitly directing attention to scene information.

To briefly illustrate the above constraints, imagine a situation in which you and your friend sit in front of the television, and your friend says “You know, I really loved those cookies at aunt Marge’s yesterday” while the television screen shows a newscaster. Both the fact that none of the referring expressions are present on the screen and the fact that the utterance refers to an event in the past, will naturally decrease reliance on information about objects and events from the television screen. Or imagine you are in the living-room, and on television you see a commercial about a cat that eats cat food. Your friend in the kitchen calls out to you “The cat has jumped on the table again”. Initial reference succeeds, and there may even be a short moment during which you attempt to establish reference from *the cat* in your friend’s utterance to the cat on the screen in front of you. However, reconciliation of the utterance with constraints provided by the current situation (e.g., the realization that your friend cannot see the television cat), presumably lead to a decrease in the relevance of the immediate scene (the television screen).

However, in the absence of such cues to bias against the strong relative importance of scene information, the Coordinated Interplay Account predicts that attended referents and associated proximal scene objects/events will play a dominant role in guiding situated comprehension processes.

Acknowledgements

This research was funded by SFB 378 “ALPHA” awarded to MWC by the German Research Foundation. We thank three anonymous reviewers for their insightful comments.

Figure Legends

Figure 1: Example images for an item in Experiment 1

Figure 2: Action sequence in Experiment 2

Figure 3: Mean inspection proportions in ms to the stereotypical and depicted agents in the unique conditions for Experiment 2

Figure 4: Mean inspection proportions in ms to the stereotypical and depicted agents in the ambiguous conditions for Experiment 2

Figure 5: Example images for an item in Experiment 3

Figure 6: The revised Coordinated Interplay Account

References

Almeida, R. G. de, Nardo, J. di, & Grunau, M. W. von. (2006). Understanding sentences in dynamic scenes. In *Proceedings of the 19th Annual CUNY Conference*. New York.

Altmann, G. T. M. (2004). Language-mediated eye-movements in the absence of a visual world: the 'blank screen paradigm'. *Cognition*, *93*, B79–B87.

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

Baayen, R. H., Pipenbrock, R., & Gulikers, L. (1995). *The celex lexical database (cd-rom)*. University of Pennsylvania, Philadelphia, PA: Linguistic Data Consortium.

Baddeley, A. D. (1986). *Working memory*. Oxford, UK; New York: Oxford University Press.

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

Chambers, C. G., Tanenhaus, M. K., & Magnuson, J. S. (2004). Actions and affordances in syntactic ambiguity resolution. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 687–696.

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

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 & F. Ferreira (Eds.), *The interface of language, vision, and action: eye movements and the visual world* (pp. 1–58). New York: Psychology Press.

Howell, D. C. (2002). *Statistical methods for psychology* (5th ed.). Pacific Grove: Duxbury.

Jackendoff, R. (2002). *Foundations of language: brain, meaning, grammar, evolution*. Oxford, UK: Oxford University Press.

Kaiser, E., & Trueswell, J. C. (2005). The role of discourse context in the processing of a flexible word-order language. *Cognition*, *94*, 113–147.

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

Knoeferle, P. (in press). Comparing the time-course of processing initially ambiguous and unambiguous

German SVO/OVS sentences in depicted events. In R. van Gompel, M. Fischer, W. Murray, & R. Hill (Eds.), *Eye movement research: insights into mind and brain.* Oxford: Elsevier.

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

Knoeferle, P., Crocker, M. W., Scheepers, C., & Pickering, M. J. (2005). The influence of the immediate visual context on incremental thematic role-assignment: Evidence from eye-movements in depicted events. *Cognition*, *95*, 95–127.

Mayberry, M., Crocker, M. W., & Knoeferle, P. (2006). A connectionist model of the coordinated interplay of scene, utterance, andworld knowledge. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society.*

Richardson, D. C., & Dale, R. (2005). Looking to understand: The coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension. *Cognitive Science*, *29*, 1045–1060.

Richardson, D. C., Dale, R., & Kirkham, N. (in press). The art of conversation is coordination: common ground and the coupling of eye movements during dialogue. *Psychological Science.*

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

Sedivy, J. C., Tanenhaus, M. K., Chambers, C. G., & Carlson, G. N. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition*, *71*, 109–148.

Snow, C. (1977). Mothers' speech research: from input to interaction. In C. Snow & C. A. Ferguson (Eds.), *Talking to children: language input and acquisition.* Cambridge, MA: Cambridge University Press.

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

Spivey, M. J., Richardson, ., & Fitneva, S. (2004). Thinking outside the brain: Spatial indices to linguistic and visual information. In J. Henderson & F. Ferreira (Eds.), *The integration of language, vision and action* (pp. 161–189). New York: Psychology Press.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*, 632–634.

Table 1

Table 1

Example item sentence set for Fig. 1 in Experiment 1

Image	Condition		Sentences
Fig. 1a	Unique identification & Depicted target	(a1)	Den Piloten verköstigt gleich der Detektiv. The pilot (PAT.) serves-food-to soon the detective. The detective will soon serve food to the pilot.
Fig. 1a	Unique identification & Stereotypical target	(a2)	Den Piloten verzaubert gleich der Zauberer. The pilot (PAT.) jinxes soon the wizard. The wizard will soon jinx the pilot.
Fig. 1a	Ambiguous identification & Depicted target	(b1)	Den Piloten bespitzelt gleich der Zauberer. The pilot (PAT.) spies-on soon the wizard. The wizard will soon spy on the pilot.
Fig. 1a	Ambiguous identification & Stereotypical Target	(b2)	Den Piloten bespitzelt gleich der Detektiv. The pilot (PAT.) spies-on soon the detective. The detective will soon spy on the pilot.
Fig. 1b	Unique identification & Depicted target	(a1')	Den Piloten bandagiert gleich der Zauberer. The pilot (PAT.) bandages soon the wizard. The wizard will soon bandage the pilot.
Fig. 1b	Unique identification & Stereotypical target	(a2')	Den Piloten bespitzelt gleich der Detektiv. The pilot (PAT.) spies-on soon the detective. The detectove will soon spy on the pilot.
Fig. 1b	Ambiguous identification & Depicted Target	(b1')	Den Piloten verzaubert gleich der Detektiv. The pilot (PAT.) jinxes soon the detective. The detective will soon jinx the pilot.
Fig. 1b	Ambiguous identification & Stereotypical. target	(b2')	Den Piloten verzaubert gleich der Zauberer. The pilot (PAT.) jinxes soon the wizard. The wizard will soon jinx the pilot.

All sentence and image materials are available from
http://www.coli.uni-saarland.de/∼knoferle/materials/jml2007.html

Table 2

Table 2

Extent of the analysis regions for Experiments 1 and 2 and their mean duration in ms per condition

Region	Extent	Mean duration per condition			
NP1	from NP1 onset to verb onset	‘The pilot’			
		(a1) 1070	(a2) 1085	(b1) 1077	(b2) 1053
VERB	from verb onset to adverb onset	‘serves-food-to’	‘jinxes’	‘spies-on’	‘spies-on’
		(a1) 1003	(a2) 1008	(b1) 989	(b2) 999
ADV	from adverb onset to NP2 onset	‘soon’			
		(a1) 1004	(a2) 1001	(b1) 1013	(b2) 1002
NP2	from NP2 onset to its offset	‘the detective’	‘the wizard’	‘the wizard’	‘the detective’
		(a1) 1056	(a2) 1042	(b1) 1062	(b2) 1064

Table 3

Table 3

Table of inspection percentages for the unique conditions in Experiment 1

Region		Condition	Target character (see Fig. 1a)				
			Backgr.	Patient	Stereotyp. Agent	Depict. Agent	Distr.
				Pilot	Wizard (‘jinx’)	Detective (serving)	
NP1	‘The pilot’	(a1) Depict. Target	41.1	44.7	7.1	5.5	1.6
		(a2) Stereotyp. Target	45.5	40.7	8.2	4.8	.9
Verb	‘serves-food-to’	(a1) Depict. Target	51.3	29.7	7.2	10.8	1.0
	‘jinxes’	(a2) Stereotyp. Target	45.7	34.9	10.3	7.4	1.7
Adv	‘soon’	(a1) Depict. Target	45.4	24.6	4.4	25.1	0.5
		(a2) Stereotyp. Target	49.2	20.9	17.5	9.0	3.4
NP2	‘the detective’	(a1) Depict. Target	43.3	21.6	9.6	24.5	0.9
	‘the wizard’	(a2) Stereotyp. Target	43.4	21.7	25.6	8.2	1.1
Late NP2	‘the detective’	(a1) Depict. Target	44.1	20.8	9.6	24.9	0.6
	‘the wizard’	(a2) Stereotyp. Target	42.2	21.1	27.7	7.5	1.5

Table 4

Table 4

Table of the results from the log-linear analyses for Experiment 1

Region	**By participants**			**By items**		
	df	$LR\chi^2(subj)$	*p*	*df*	$LR\chi^2(item)$	*p*
Log-linear analyses for the unique identification conditions						
ADV						
Target character	1	*n.s.*	*n.s.*	1	5.28	0.02*
Target character x Target type	1	24.06	0.000*	1	*n.s.*	*n.s.*
NP2						
Target character	1	0.01	0.9	1	0.1	0.9
Target character x Target type	1	27.15	0.000*	1	42.75	0.000*
Late NP2						
Target character	1	1.61	0.21	1	1.61	0.21
Target character x Target type	1	70.36	0.000*	1	64.55	0.000*
Log-linear analyses for the ambiguous identification conditions						
ADV						
Target character	1	11.79	0.001*	1	11.79	0.001*
Target character x Target type	1	1.34	0.25	1	1.64	0.20
NP2						
Target character	1	*n.s.*	*n.s.*	1	5.24	0.02*
Target character x Target type	1	*n.s.*	*n.s.*	1	20.33	0.000*
Late NP2						
Target character	1	4.76	0.03*	1	4.76	0.03*
Target character x Target type	1	41.28	0.000*	1	32.75	0.000*

*$p < 0.05$; *n.s.* indicates that the k-way likelihood ratio effect was non-significant, showing that the contribution of k-order effects to the model was not significant. Note that a non-significant likelihood ratio effect takes precedence over significant partial associations (see, e.g., Howell, 2002).

Table 5

Table 5

Table of inspection percentages for the ambiguous conditions in Experiment 1

Region		Condition	Target character (see Fig. 1a)				
			Backgr.	Patient	Stereotyp. Agent	Depict. Agent	Distr.
				Pilot	Detective (‘spies-on’)	Wizard (spying)	
NP1	‘The pilot’	(b1) Depict. Target	43.7	40.2	7.1	7.1	2.0
		(b2) Stereotyp. Target	49.8	36.5	6.0	6.9	.9
Verb	‘spies-on’	(b1) Depict. Target	48.5	35.2	6.6	9.2	.5
		(b2) Stereotyp. Target	41.6	39.9	8.4	10.1	0
Adv	‘soon’	(b1) Depict. Target	41.7	27.0	10.0	20.9	0.4
		(b2) Stereotyp. Target	44.2	25.4	11.7	18.8	0
NP2	‘the wizard’	(b1) Depict. Target	43.9	23.9	8.0	24.1	0.2
	‘the detective’	(b2) Stereotyp. Target	43.1	25.4	19.4	10.8	1.3

Table 6

Table 6

Table of inspection percentages for the unique conditions in Experiment 2

Region		Condition	Target character (see Fig. 2a)				
			Backgr.	Patient	Stereotyp. Agent	Depict. Agent	Distr.
				Pilot	Wizard (‘jinx’)	Detective (serving)	
NP1	‘The pilot’	(a1) Depict. Target	24.2	42.6	16.2	12.8	4.3
		(a2) Stereotyp. Target	24.6	41.4	13.3	16.2	4.3
Verb	‘serves-food-to’	(a1) Depict. Target	28.5	26.8	18.4	24.3	2.1
	‘jinxes’	(a2) Stereotyp. Target	28.2	27.1	24.7	17.6	2.4
Adv	‘soon’	(a1) Depict. Target	31.5	20.4	20.4	26.2	1.5
		(a2) Stereotyp. Target	30.1	23.8	26.2	17.6	2.3
VerbAdv	‘serves-food-to soon’	(a1) Depict. Target	29.5	23.8	19.6	25.5	1.7
	‘jinxes soon’	(a2) Stereotyp. Target	29.5	25.9	26.2	16.9	1.6
NP2	‘the detective’	(a1) Depict. Target	31.1	16.4	16.7	34.8	1.0
	‘the wizard’	(a2) Stereotyp. Target	27. 7	15.8	36.5	19.6	0.4

Table 7

Table 7

Table of the results from the log-linear analyses for Experiment 2

Region	**By participants**			**By items**		
	df	$LR\chi^2(subj)$	*p*	*df*	$LR\chi^2(item)$	*p*
Log-linear analyses for the unique identification conditions						
NP1						
Target character	1	0.12	0.73	1	0.12	0.73
Target character x Target type	1	2.84	0.09	1	4.22	0.04*
VERB						
Target character	1	0.04	0.83	1	0.04	0.84
Target character x Target type	1	4.41	0.04*	1	2.23	0.14
ADV						
Target character	1	0.36	0.55	1	0.36	0.55
Target character x Target type	1	3.79	0.05	1	7.60	0.01*
Verbadv						
Target character	1	0.77	0.38	1	0.77	0.38
Target character x Target type	1	11.74	0.001*	1	11.76	0.001*
NP2						
Target character	1	0.50	0.48	1	0.50	0.48
Target character x Target type	1	33.88	0.000*	1	27.39	0.000*
Log-linear analyses for the ambiguous identification conditions						
NP1						
Target character	1	1.71	0.19	1	1.71	0.19
Target character x Target type	1	4.28	0.04*	1	4.33	0.04*
VERB						
Target character	1	2.04	0.15	1	2.04	0.15
Target character x Target type	1	0.00	0.97	1	0.45	0.50
ADV						
Target character	1	0.40	0.53	1	0.40	0.53
Target character x Target type	1	0.24	0.63	1	0.26	0.61
Verbadv						
Target character	1	3.32	0.07	1	3.32	0.07
Target character x Target type	1	0.00	1.00	1	0.11	0.74
NP2						
Target character	1	0.76	0.38	1	0.76	0.38
Target character x Target type	1	21.79	0.000*	1	21.12	0.000*
Three-way interaction of target character, identification and target type						
Verbadv						
Target character x Identification x Target type	1	13.57	0.001*	1	*n.s.*	*n.s.*

*$p < 0.05$; *n.s.* indicates that the k-way likelihood ratio effect was non-significant. Note that a non-significant likelihood ratio effect takes precedence over significant partial associations (see, e.g., Howell, 2002).

Table 8

Table 8

Table of inspection percentages for the ambiguous conditions in Experiment 2

Region		Condition	Target character (see Fig. 2a)				
			Backgr.	Patient	Stererotyp. Agent	Depict. Agent	Distr.
				Pilot	Detective (‘spies-on’)	Wizard (spying)	
NP1	‘The pilot’	(b1) Depict. Target	26.8	43.4	13.0	11.6	5.2
		(b2) Stererotyp. Target	24.1	44.3	10.8	17.2	3.6
Verb	‘spies-on’	(b1) Depict. Target	23.5	29.6	23.0	21.0	2.9
		(b2) Stererotyp. Target	27.7	23.3	26.1	20.2	2.8
Adv	‘soon’	(b1) Depict. Target	30.4	20.9	26.6	21.3	0.8
		(b2) Stererotyp. Target	24.9	17.7	27.7	28.5	1.2
VerbAdv	‘spies-on soon’	(b1) Depict. Target	26.7	25.1	24.9	21.3	2.0
	‘jinxes soon’	(b2) Stererotyp. Target	26.4	21.5	27.5	22.6	2.0
NP2	‘the wizard’	(b1) Depict. Target	25.9	14.5	20.7	38.6	0.3
	‘the detective’	(b2) Stererotyp. Target	31.9	16.7	31.2	19.1	1.1

Table 9

Table 9

Example item sentence set for Fig. 5 in Experiment 3

Image	Condition		Sentences
Fig. 5a-c	Future tense	(a)	Der Kellner poliert demnächst die Kristallgläser.
			The waiter (subject/agent) polishes soon the crystal glasses (object/patient).
			The waiter will soon polish the crystal glasses.
Fig. 5a-c	Past tense	(b)	Der Kellner polierte kürzlich die Kerzenständer.
			The waiter (subject/agent) polished recently the chandeliers (object/patient).
			The waiter recently polished the chandeliers.
Fig. 5a'-c'	Future tense	(a')	Der Kellner poliert demnächst die Kerzenständer.
			The waiter (subject/agent)polishes soon the chandeliers (object/patient).
			The waiter will soon polish the chandeliers.
Fig. 5a'-c'	Past tense	(b')	Der Kellner polierte kürzlich die Kristallgläser.
			The waiter (subject/agent) polished the crystal glasses (object/patient).
			The waiter recently polished the crystal glasses.

All sentence and image materials are available from http://www.coli.uni-saarland.de/∼knoferle/materials/jml2007.html

Table 10

Table 10

Extent of the analysis regions for Experiment 3 and their mean duration in ms per condition

Region	Extent	Mean duration per condition	
NP1	from NP1 onset to verb onset	‘The waiter’	
		(a) 1316	(b) 1315
VERB	from verb onset to adverb onset	‘polishes’	‘polished’
		(a) 892	(b) 909
ADV	from adverb onset to NP2 onset	‘soon’	‘recently’
		(a) 1298	(b) 1240
NP2	from NP2 onset to its offset	‘the crystal glasses’	‘the chandeliers’
		(a) 892	(b) 853

Table 11

Table 11

Table of inspection percentages for Experiment 3

Region		Condition	Target objects (see Fig. 5c)				
			Background	Agent	Past target	Future target	Distractors
				waiter	chandeliers	crystal glasses	
NP1	'The waiter'	(a) Future tense	27.9	32.0	16.9	13.6	9.6
		(b) Past tense	28.0	31.3	16.6	14.1	10.0
Verb	'polishes'	(a) Future tense	24.2	21.0	35.2	14.0	5.7
	'polished'	(b) Past tense	27.1	19.5	31.1	15.9	6.5
Adv	'soon'	(a) Future tense	27.7	20.0	24.5	24.6	3.1
	'recently'	(b) Past tense	25.9	18.2	32.7	18.7	4.5
NP2	'crystal glasses'	(a) Future tense	25.6	13.4	21.1	36.1	3.7
	'chandeliers'	(b) Past tense	26.3	18.7	32.0	18.7	4.3
Late NP2	'crystal glasses'	(a) Future tense	23.3	9.8	22.2	42.7	2.0
	'chandeliers'	(b) Past tense	27.0	16.8	32.3	17.1	6.8

Table 12

Table of the results from the log-linear analyses for Experiment 3

Region	**By participants**			**By items**		
	df	$LR\chi^2(subj)$	*p*	*df*	$LR\chi^2(item)$	*p*
VERB						
Target object	1	80.32	0.000*	1	80.32	0.000*
Target object x Tense	1	1.30	0.26	1	2.46	0.12
ADV						
Target object	1	12.06	0.001*	1	12.06	0.001*
Target object x Tense	1	*n.s.*	*n.s.*	1	*n.s.*	*n.s.*
Late NP2						
Target object	1	1.48	0.22	1	1.48	0.22
Target object x Tense	1	35.78	0.000*	1	36.75	0.000*

*$p < 0.05$; *n.s.* indicates that the k-way likelihood ratio effect was non-significant, showing that the contribution of k-order effects to the model was not significant. Note that a non-significant likelihood ratio effect takes precedence over significant partial associations (see, e.g., Howell, 2002).

Figure 1

Click here to download high resolution image

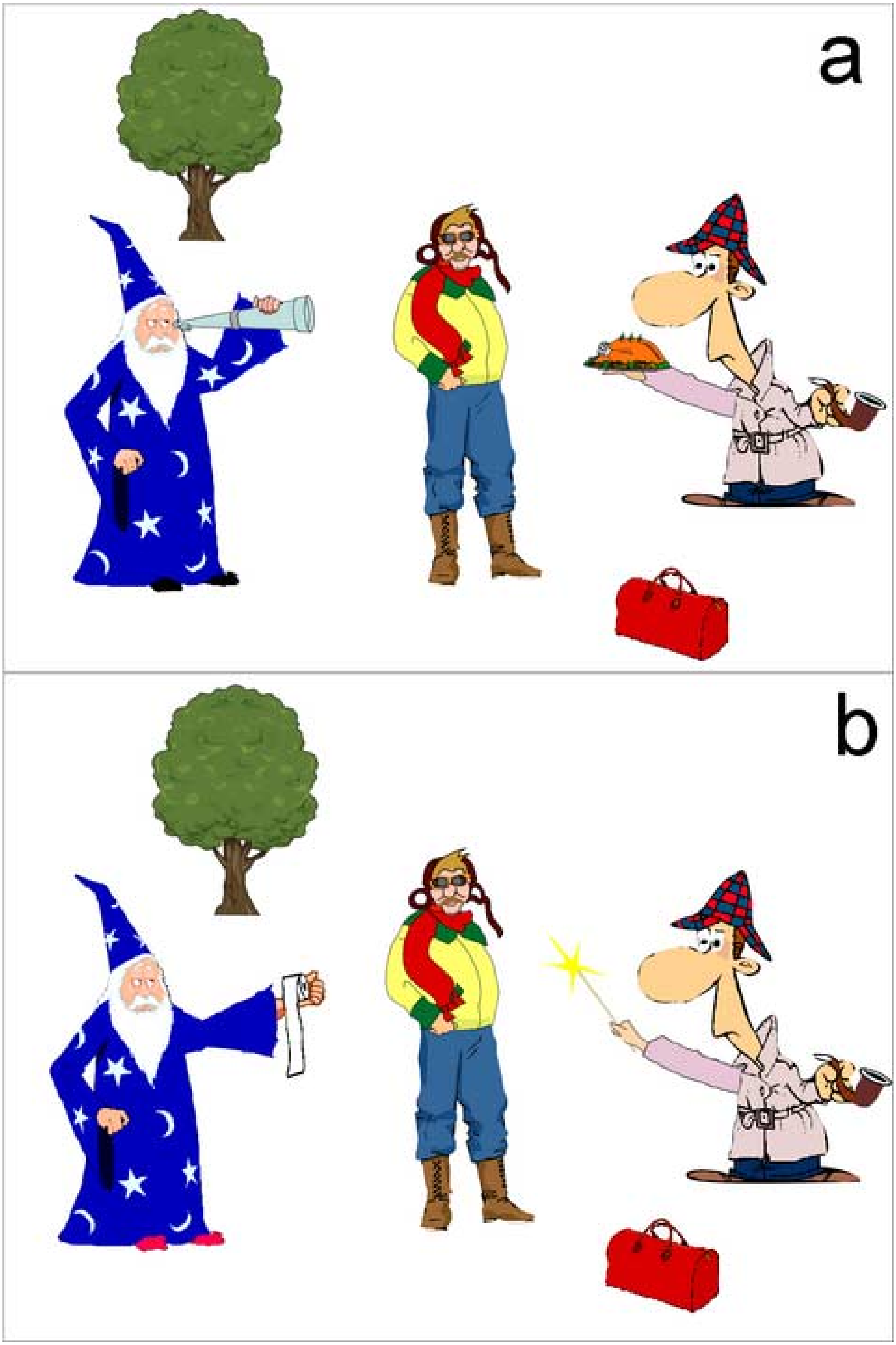


Figure 2

Click here to download high resolution image

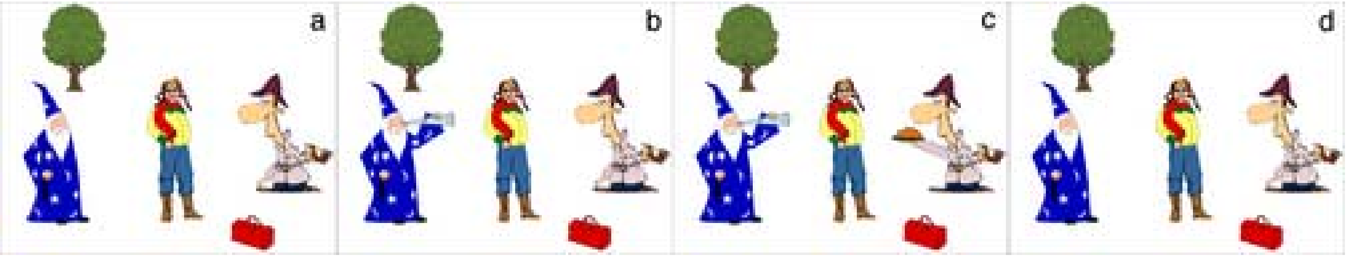


Figure 3

Click here to download high resolution image

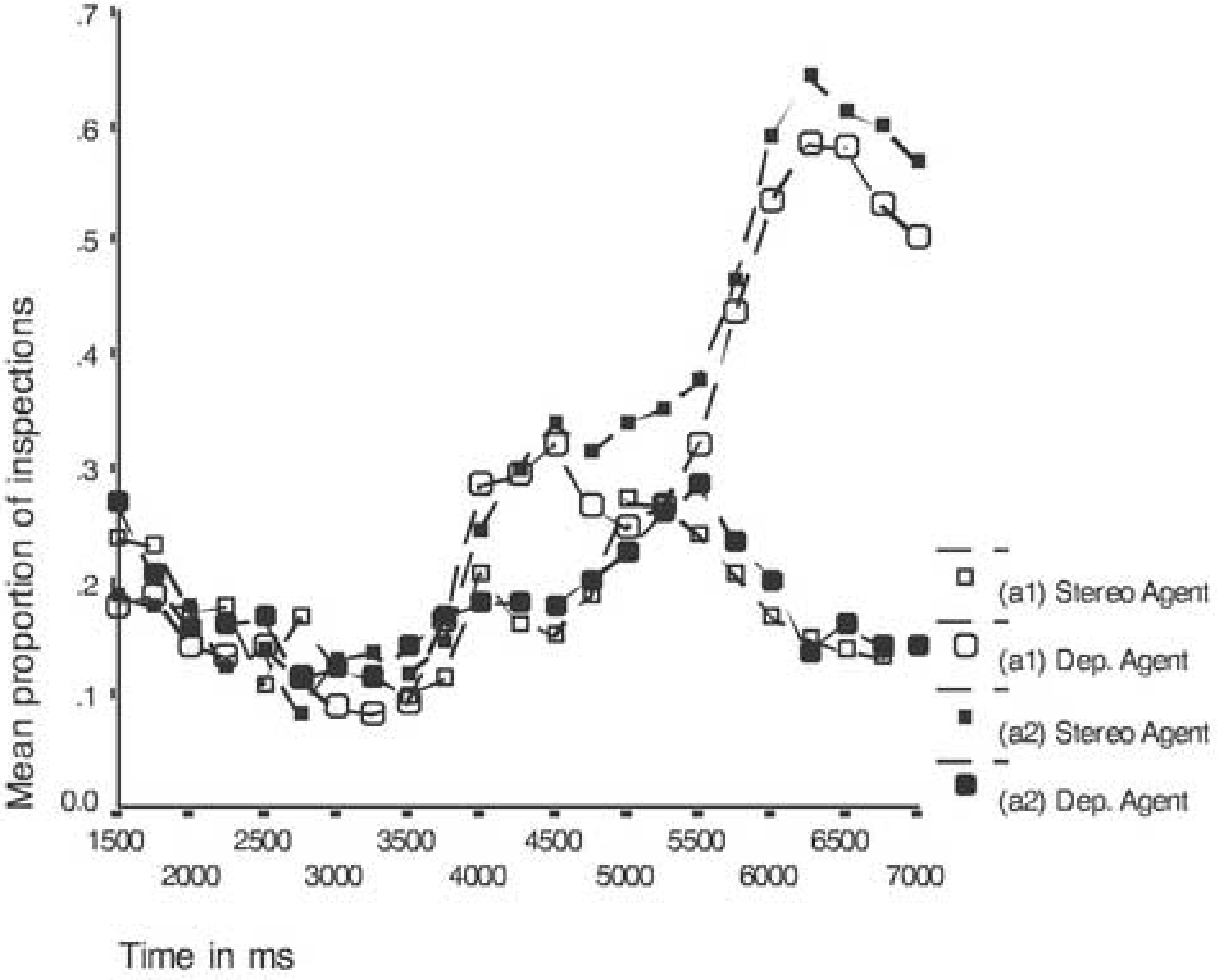


Figure 4

Click here to download high resolution image

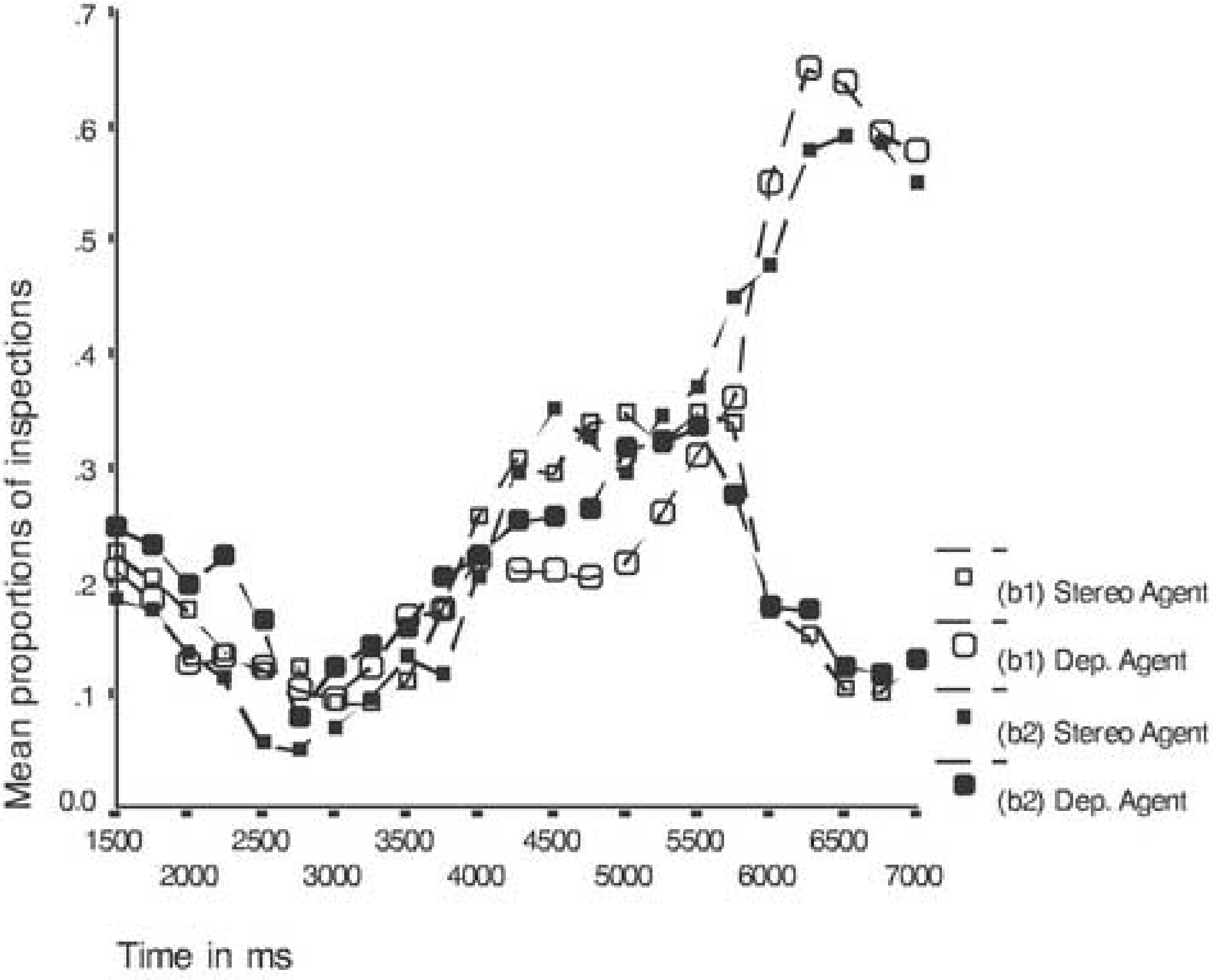


Figure 5
Click here to download high resolution image

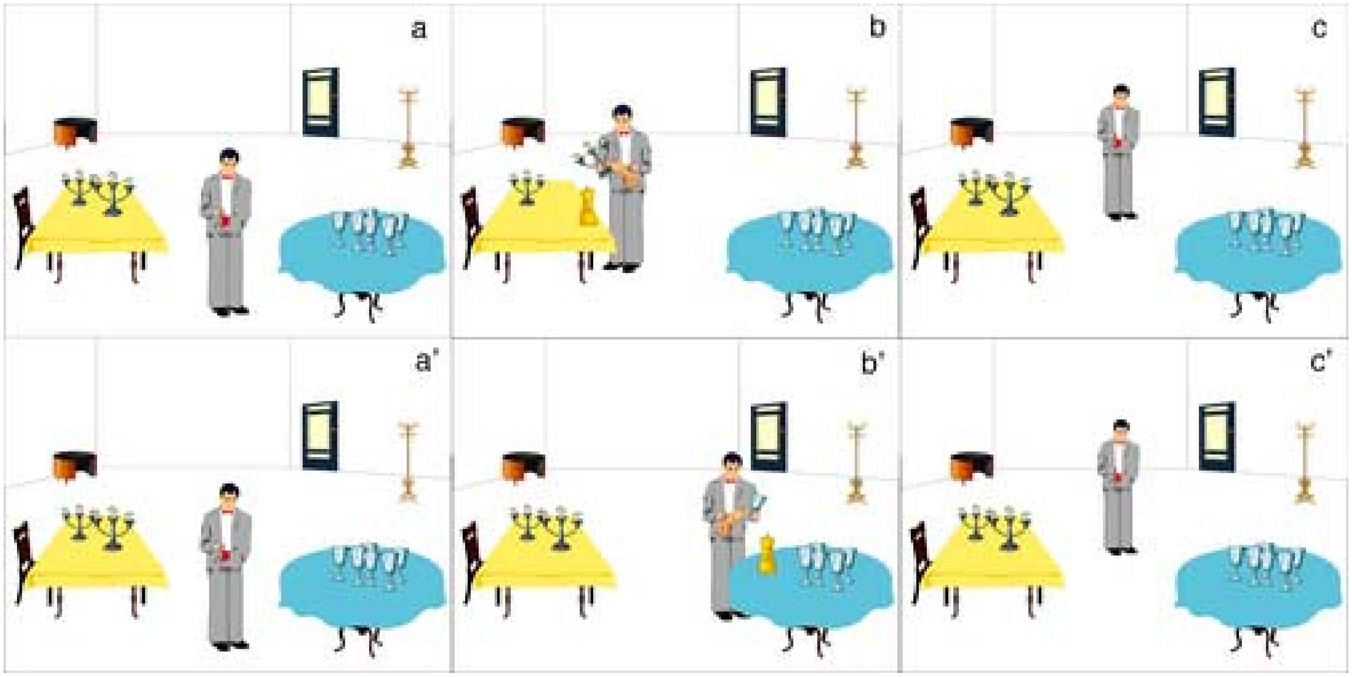
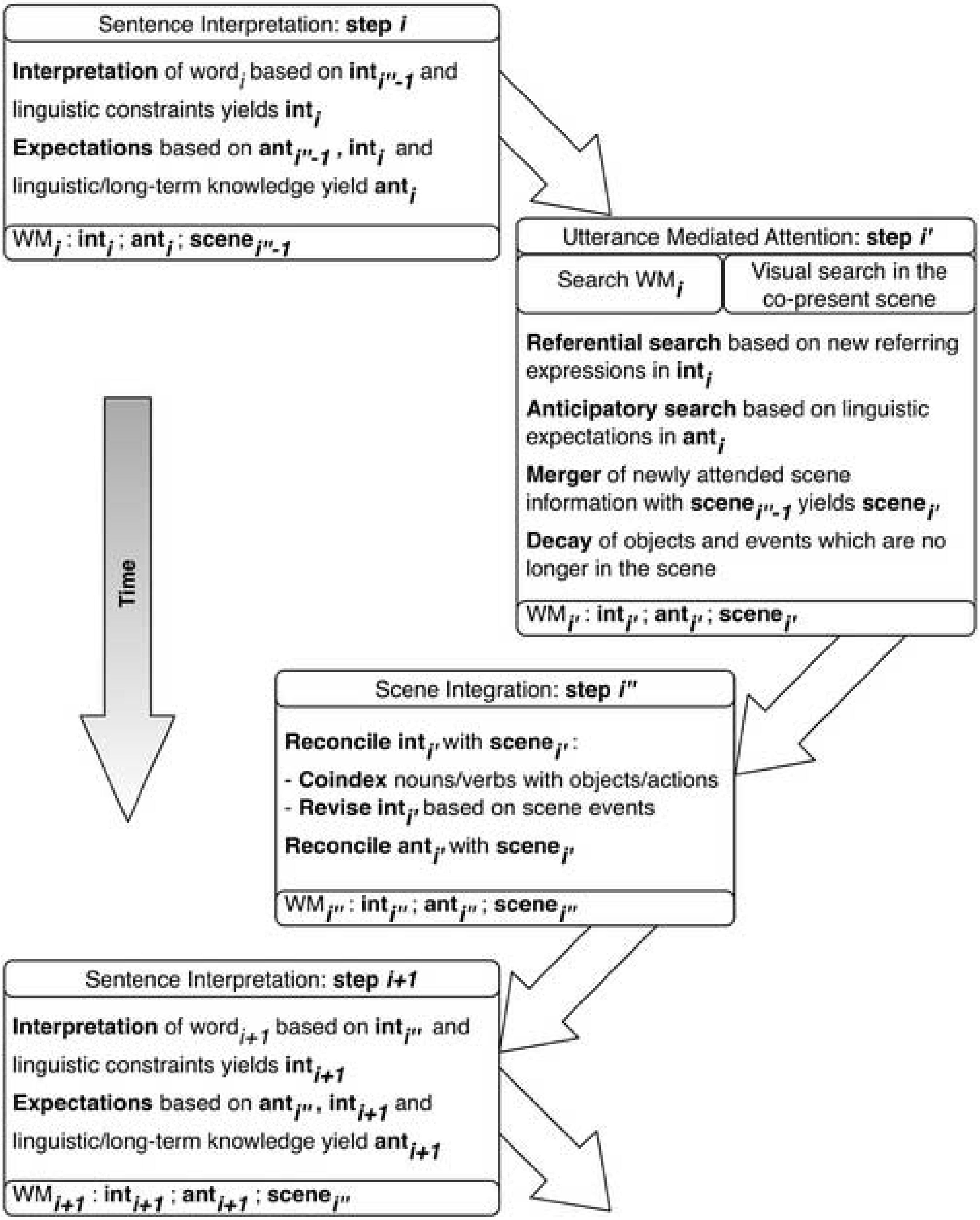


Figure 6

Click here to download high resolution image


Sentence Interpretation: step i
Interpretation of word$_i$ based on int$_{i''-1}$ and linguistic constraints yields int$_i$
Expectations based on ant$_{i''-1}$, int$_i$ and linguistic/long-term knowledge yield ant$_i$
WM$_i$: int$_i$; ant$_i$; scene$_{i''-1}$
Utterance Mediated Attention: step i'
Search WM$_i$
Visual search in the co-present scene
Referential search based on new referring expressions in int$_i$
Anticipatory search based on linguistic expectations in ant$_i$
Merger of newly attended scene information with scene$_{i''-1}$ yields scene$_{i'}$
Decay of objects and events which are no longer in the scene
WM$_{i'}$: int$_{i'}$; ant$_{i'}$; scene$_{i'}$
Time
Scene Integration: step i''
Reconcile int$_{i'}$ with scene$_{i'}$:
- Coindex nouns/verbs with objects/actions
- Revise int$_{i'}$ based on scene events
Reconcile ant$_{i'}$ with scene$_{i'}$
WM$_{i''}$: int$_{i''}$; ant$_{i''}$; scene$_{i''}$
Sentence Interpretation: step $i+1$
Interpretation of word$_{i+1}$ based on int$_{i''}$ and linguistic constraints yields int$_{i+1}$
Expectations based on ant$_{i''}$, int$_{i+1}$ and linguistic/long-term knowledge yield ant$_{i+1}$
WM$_{i+1}$: int$_{i+1}$; ant$_{i+1}$; scene$_{i''}$