Cognitive Science 30 (2006) 481–529

The Coordinated Interplay of Scene, Utterance, and World Knowledge: Evidence From Eye Tracking

Pia Knoeferle, Matthew W. Crocker

Computational Linguistics, Saarland University, Germany

Received 11 December 2004; received in revised form 19 October 2005; accepted 17 January 2006

Abstract

Two studies investigated the interaction between utterance and scene processing by monitoring eye movements in agent–action–patient events, while participants listened to related utterances. The aim of Experiment 1 was to determine if and when depicted events are used for thematic role assignment and structural disambiguation of temporarily ambiguous English sentences. Shortly after the verb identified relevant depicted actions, eye movements in the event scenes revealed disambiguation. Experiment 2 investigated the relative importance of linguistic/world knowledge and scene information. When the verb identified either only the stereotypical agent of a (nondepicted) action, or the (nonstereotypical) agent of a depicted action as relevant, verb-based thematic knowledge and depicted action each rapidly influenced comprehension. In contrast, when the verb identified both of these agents as relevant, the gaze pattern suggested a preferred reliance of comprehension on depicted events over stereotypical thematic knowledge for thematic interpretation. We relate our findings to language comprehension and acquisition theories.

Keywords: Situated comprehension; Eye tracking; Processing account; Depicted events

1. Introduction

For the comprehension of spoken utterances that relate to a visual scene, listeners can draw on information provided by entities and events in the immediate scene, as well as linguistic, semantic, and world knowledge. The monitoring of eye movements in scenes has been successfully exploited for the investigation of a variety of psycholinguistic research questions in online spoken language comprehension. Among these are the effects of discourse context (Kaiser & Trueswell, 2005), incremental thematic interpretation (Kamide, Altmann, & Haywood, 2003; Kamide, Scheepers, & Altmann, 2003), binding theory (Runner, Sussman, & Tanenhaus, 2003), filler-gap dependencies (Sussman & Sedivy, 2003), and child sentence comprehension (Snedeker & Trueswell, 2004; Trueswell, Sekerina, Hill, & Logrip, 1999).

Correspondence should be addressed to Pia Knoeferle, Department of Computational Linguistics, Building C71, Room 1.20, Postbox 15 11 50, Saarland University, 66041 Saarbrücken, Germany. E-mail: knoeferle@coli.uni-sb.de

However, comparatively little attention has been paid to the nature of the online interaction among utterance comprehension mechanisms, linguistic and world knowledge, and scene processing itself. Eye movements in scenes during utterance comprehension have revealed that attention in visual scenes is closely time-locked with real-time comprehension (Cooper, 1974; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). They have further revealed that the utterance can direct attention to an object in the scene even before that object is directly referred to by the utterance (Altmann & Kamide, 1999). Moreover, we know that diverse informational sources—linguistic and world knowledge (e.g., Chambers, Tanenhaus, Filip, & Carlson, 2002; Kamide, Scheepers, & Altmann, 2003) as well as scene information (e.g., Knoeferle, Crocker, Scheepers, & Pickering, 2005; Sedivy, Tanenhaus, Chambers, & Carlson, 1999; Spivey, Tanenhaus, Eberhard, & Sedivy, 2002; Tanenhaus et al., 1995)—can each rapidly influence online utterance comprehension and structural disambiguation (see also Altmann, 2004; Spivey & Geng, 2001).

These findings allow us to identify two fundamental dimensions of the interaction among scene, utterance, and linguistic and world knowledge. One dimension is the temporal coordination between utterance and scene processing (henceforth referred to as *temporal dimension*). The second dimension is the rapid influence of diverse informational sources on incremental utterance comprehension (henceforth dubbed *informational dimension*).

We carried out two eye movement experiments to further explore these two dimensions of situated utterance comprehension. First, Tanenhaus et al. (1995) showed a close time-lock between utterance comprehension and attention in the scene. We extend their findings by examining the precise time-course with which scene information informs online utterance comprehension and disambiguation as a function of when the utterance identifies that scene information as relevant for comprehension (Experiment 1). Further, we add to prior research that has shown the influence of verb-based world knowledge about likely role fillers (e.g., Kamide, Scheepers, & Altmann, 2003), as well as of verb-mediated depicted actions and their associated role relations (Knoeferle et al., 2005). Specifically, we investigate the relative importance of these two informational sources—how important immediately depicted events are relative to world knowledge about plausible role fillers for incremental thematic interpretation (Experiment 2). We suggest that a better understanding of these two processing dimensions is essential to our understanding of how full utterance interpretation results from the online reconciliation of linguistic/world knowledge and immediate scene information.

An important initial finding relevant to both the coordination of utterance comprehension with scene processing and the rapid influence of scene information was made by Tanenhaus et al. (1995). In instructions such as *Put the apple on the towel in the box,* the phrase *on the towel* can be temporarily analyzed as a modifier of the first noun phrase, identifying the location of the apple, or as an argument of the verb, identifying where to put the apple. In a context containing one apple on a towel (location) and an empty towel (destination), people inspected the apple when hearing *Put the apple*. Having heard *on the towel,* listeners mostly fixated the empty towel, suggesting interpretation of the ambiguous phrase as destination. In a scene with two apples (of which one was on a towel) and an empty towel, fixation patterns differed from the start of the utterance in comparison with the one-apple context. People's eye movements oscillated between the two apples during *the apple* and then settled on the apple on the towel, indicating interpretation of the phrase *on the towel* as the location of the apple.

The core findings are that utterance interpretation directs attention in the scene (see also Cooper, 1974; Spivey, Tyler, Eberhard, & Tanenhaus, 2001), and that the visual referential context rapidly influences the incremental structuring of the utterance. The rapidity of scene influence was confirmed by the fact that eye movements differed between the two visual context conditions from the onset of the utterance (see also Spivey et al., 2002).

What the fixation patterns in the studies by Tanenhaus et al. (1995) do not, however, permit us to determine is whether scene information influenced structuring and interpretation of the utterance in a manner closely time-locked to, or independently of, when the utterance identified that scene information as relevant. In one interpretation, comprehension of the utterance (i.e., *the apple*) directed attention in the scene, and this triggered the construction and use of the appropriate referential context (one apple, two apples). A second possible interpretation is that people acquired the referential context temporally independently of—perhaps even prior to—hearing the utterance, and then accessed that context much as they would access a prior discourse context (e.g., Altmann & Steedman, 1988). Findings by Tanenhaus and colleagues are compatible with both interpretations. The fact that eye movement patterns differed from the start of the utterance between the two contexts makes it impossible to determine precisely whether or not identification of relevant scene information by the utterance was necessary for that scene information to affect structural disambiguation.

Studies by Knoeferle et al. (2005) and Sedivy et al. (1999) further extend insights on the informational and temporal processing dimensions. Both confirm the rapid influence of scene information, and appear to suggest that the use of scene information is triggered when the utterance identifies it as relevant. Studies by Sedivy et al. demonstrated that the time course of establishing reference to objects in a scene depended on whether there was referential contrast between two same-category objects (two glasses) or not (Sedivy et al., 1999). In the referential contrast condition, an example scene contained two tall objects (a glass and a pitcher), a small glass, and a key. In the no-referential-contrast condition, the scene contained two tall objects (a glass and pitcher), a key, and a file folder, but no contrasting object of the category "glass." Sedivy et al. (1999) found no differences in the gaze pattern between the two context types when participants heard *Pick up the*. Only after people had heard *Pick up the tall* did they look more quickly at the target referent (the tall glass) than at the other tall object (a pitcher) when the visual context displayed a contrasting object of the same category as the target (a small glass) than when it did not.

Knoeferle et al. (2005) found a rapid effect of depicted events showing who did what to whom on incremental thematic role assignment and structural disambiguation of German sentences. Their studies investigated comprehension of subject–verb–object (SVO)/object–verb–subject (OVS) sentences. Although both of these orders are grammatical, the subject-initial order is preferred (e.g., Hemforth, 1993). A case-marked article can (but does not always) determine the grammatical function and thematic role of the noun phrase it modifies. Knoeferle et al. (2005) investigated SVO/OVS sentences when neither case-marking nor other cues in the utterance determined the correct syntactic and thematic relations prior to the sentence-final accusative (SVO) or nominative (OVS) case-marked noun phrase. For early disambiguation, listeners had to rely on depicted event scenes that showed a princess washing a pirate, while a fencer painted that princess. Listeners heard *Die Prinzessin wäscht/malt den Pirat/der Fechter* ('The princess (amb.) washes/paints the pirate (ACC)/ the fencer (NOM)').

The verb in the utterance identified either the washing or the painting action, establishing the princess as event agent (SVO) or patient (OVS), respectively. Eye movements for the SVO in comparison with the OVS condition did not differ prior to the verb. After the verb had been encountered, eye movements to the patient (the pirate) and the agent (the fencer) of the action for SVO and OVS, respectively, revealed rapid thematic role assignment and structural disambiguation. Gaze patterns that are suggestive of such close temporal coordination were also observed in a further experiment by Knoeferle and colleagues when not the verb but subtle event cues such as temporal adverbs made depicted events available for early thematic role assignment and structural disambiguation.

Although not directly aimed at investigating the utterance-mediated influence of the scene, findings by Sedivy et al. (1999) and Knoeferle et al. (2005) provide a clearer picture than gaze patterns in Tanenhaus et al. (1995) of when—in relation to its identification as relevant by the utterance—scene information influences comprehension. Eye movements to the target object (the tall glass or a patient or agent in the scene) did not differ between the experimental conditions (referential contrast/no contrast and SVO/OVS) prior to the onset of the critical word that identified relevant scene information. (Sedivy et al., 1999, however, provided no inferential analyses for the time prior to adjective onset.) Only after people had heard the adjective *tall* (Sedivy et al., 1999), or the verb (Knoeferle et al., 2005) that identified relevant scene information, did that scene information (contrast between a small or tall glass in Sedivy et al., 1999, and scene actions in Knoeferle et al., 2005) appear to influence comprehension. These gaze patterns suggest a tight coordination between when the utterance identified scene information as relevant and when that scene information was used for online comprehension. The preceding studies further demonstrate that scene information directly and rapidly determined the preferred structuring (Knoeferle et al., 2005; Tanenhaus et al., 1995) and semantic interpretation (Sedivy et al., 1999) of an utterance. They provide strong support for the fundamental importance of the visual context in the interpretation and structuring of an utterance.

Other experiments, in contrast, provide evidence for the importance of linguistic and world knowledge during utterance comprehension in visual scenes (e.g., Altmann & Kamide, 1999; Chambers et al., 2002; Kamide, Scheepers, & Altmann, 2003). Kamide, Scheepers, and Altmann (2003) demonstrated the incremental use of syntactic and verb-based thematic role knowledge during utterance comprehension as revealed by anticipatory eye movements. Participants inspected images showing a hare, a cabbage, a fox, and a distractor object while hearing SVO/OVS sentences such as *Der Hase frisst gleich den Kohl* (‘The hare (NOM) eats soon the cabbage (ACC)’) and *Den Hasen frisst gleich der Fuchs* (‘The hare (ACC) eats soon the fox (NOM)’). The nominative (subject) and accusative (object) case-marking on the article of the first noun phrase together with world knowledge about what is likely to eat what extracted at the verb allowed anticipation of the correct postverbal referent. This was evidenced by an increase in anticipatory eye movements to the cabbage after participants had heard ‘The hare (NOM) eats … ’ and to the fox after having encountered ‘The hare (ACC) eats … .’

In sum, findings concerning the temporal dimension suggest a tight coordination between utterance identification of scene information as relevant, and the use of that scene information for comprehension. It is further clear that depicted events and linguistic/world knowledge can each rapidly influence online thematic role assignment (informational dimension). What remains open is the relative importance of these two informational sources for utterance compre-

hension. We suggest that a temporal coordination mechanism of utterance comprehension with attention in the scene might involve a strategy that prioritizes checking the scene for relevant objects and events rather than solely relying on linguistic and world knowledge.

Most existing psycholinguistic theories or frameworks provide a detailed account of the time course and the kinds of linguistic and world knowledge that influence online comprehension (e.g., Frazier & Clifton, 1996; MacDonald, Pearlmutter, & Seidenberg, 1994; Pickering, Traxler, & Crocker, 2000; Townsend & Bever, 2001; Trueswell & Tanenhaus, 1994). Because they have been developed based on insights from reading research, they do not explicitly include scene information. As a result, deriving predictions on the time course with which scene information is used in situated utterance comprehension from them is not straightforward. Current frameworks for situated language processing, in contrast, explicitly include scene and language information. They are, however not yet sufficiently developed to make clear predictions concerning the temporal coordination of scene and utterance processing, and the relative importance of diverse informational sources during comprehension (e.g., Barsalou, 1999a, 1999b; Bergen & Chang, 2005; Bergen, Chang, & Narayan, 2004; Chambers, Tanenhaus, & Magnuson, 2004; Jackendoff, 2002; Zwaan, 2004).

In contrast with adult sentence comprehension theories, language development research has paid greater consideration to the coordination of scene and utterance processing, as well as to the importance of diverse informational sources. Experimental findings suggest that the influence of nonlinguistic information (e.g., objects and events) on language acquisition is highly dependent on the time-lock between a child's attention to, and child-directed speech about, these objects and events (e.g., Dunham, Dunham, & Curwin, 1993; Harris, Jones, Brookes, & Grant, 1986). With respect to the importance of diverse informational sources, the acquisition accounts by Gleitman (1990), and Gillette, Gleitman, Gleitman, and Lederer (1999) emphasize the importance of linguistic knowledge for child language acquisition. At the same time, they acknowledge that no one informational source can account for full language acquisition:

> The position I have been urging is that children usually succeed in ferreting out the forms and the meaning of the language just because they can play off these two imperfect and insufficient data bases (the saliently interpretable events and the syntactically interpreted utterances) against each other to derive the best fit between them. (Gleitman, 1990, p. 50; see also Gillette et al., 1999)

Gillette et al. (1999) suggested further that the presence of objects and events is fundamental for the acquisition of concrete nouns and verbs, thus highlighting the importance of the immediate scene: "word-to-world pairing … must be the rockbottom foundation for vocabulary acquisition" (p. 169). Snow (1977) suggested that early language is largely about objects and events in the immediate environment of a child, thus corroborating this proposal. For the acquisition of more abstract terms, in contrast, Gillette et al. (1999) suggested a lesser importance of information from the immediate scene.

The preceding developmental research emphasizes that a close temporal coordination of language about and attention to objects and events is fundamental to the influence of scene information on language acquisition. They further suggest a great—albeit not necessarily primary—importance of the immediate scene for the acquisition of concrete vocabulary. The two experiments presented here investigate whether these acquisition mechanisms have, at least to some extent, fundamentally shaped adult comprehension mechanisms.

If this were the case, then we might expect to find a tight temporal coordination between when a concrete word in the utterance directs a listener's attention to an object or event, and when the attended scene information influences comprehension. In contrast, when more abstract (e.g., tense) cues identify an object or event as relevant, we would expect a less robust coordination between utterance and scene processing. A language system that preserves at least some mechanisms of acquisition might further be geared toward acquiring new information from a scene rather than relying solely on linguistic and world knowledge. If this were the case then we would expect to find a greater relative importance of an immediately depicted, verb-mediated event (Knoeferle et al., 2005) over long-term knowledge of stereotypical events (Kamide, Scheepers, & Altmann, 2003).

Experiment 1 examined the temporal dimension. It investigated whether the influence of depicted events on incremental thematic role assignment and structural disambiguation are closely temporally coordinated with their identification as relevant by the utterance. Experiment 1a examined such close temporal coordination when the actions and their associated role relations were identified as relevant by an action verb. In addition we included more indirect (tense) cues to investigate whether a temporally coordinated mediation of depicted events applies even without direct verb-action reference (Experiment 1b). Furthermore, we aim to see whether the influence of depicted events on structural disambiguation in German SVO/OVS sentences extends to the English main clause (MV)/reduced relative (RR) ambiguity, and hence generalizes to another language and construction.

Experiment 2 investigated the informational dimension, and directly compared the importance of verb-mediated world knowledge about likely role fillers (Kamide, Scheepers, & Altmann, 2003) with that of verb-mediated depicted events (Knoeferle et al., 2005). A first condition pair was designed to replicate existing findings and to ensure that depicted events and verb-based stereotypical role knowledge are comparatively effective and can each rapidly influence thematic role assignment. The verb in the utterance either identified a scene agent as relevant on the basis of its associated stereotypical thematic role knowledge (see Kamide, Scheepers, & Altmann, 2003), or it identified a (different) agent as the likely target by means of the action that agent performed in the scene (Knoeferle et al., 2005). Second, to determine the relative importance of depicted events and verb-based thematic role knowledge, we designed a condition pair for which one and the same verb identified two different scene characters as relevant at the same time. One character was identified as relevant by verb-based thematic role knowledge (rather than immediate event depictions), and another character through the depicted event it performed rather than verb-based stereotypical knowledge. Here, stereotypical thematic role knowledge and depicted events each suggested a different scene character as the appropriate target, forcing the comprehension system to choose between two sources of information in anticipating the most likely agent role filler.

2. Experiment 1

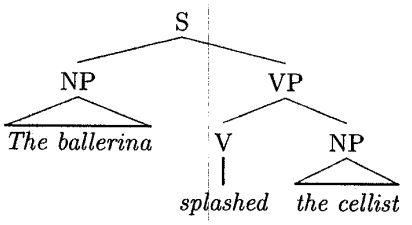
Experiment 1 examines whether utterance-directed attention in the scene and the use of scene information for structural disambiguation and incremental thematic role assignment are closely temporally coordinated. To examine this we used the English MV/RR ambiguity. Ex-

amples 1 and 2 illustrate this type of ambiguity (for a complete example of an item sentence set see Table 1).

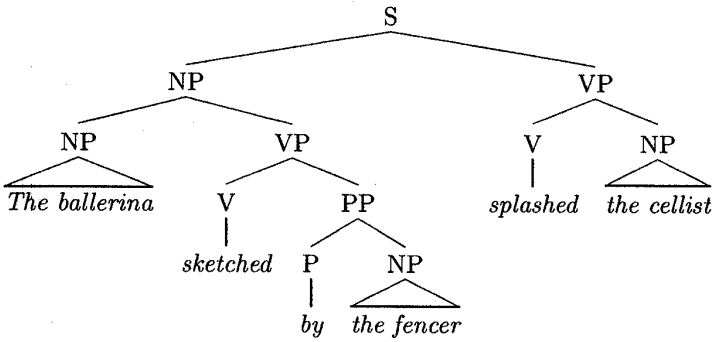
The ballerina splashed the cellist. (1)

The ballerina sketched by the fencer splashed the cellist. (2)

Examples (1) and (2) are ambiguous up to the second argument in two respects. First, they are structurally ambiguous. The verb phrase can be directly attached to the S node as the main predicate of the sentence (main clause analysis, see Example (3)), which is the correct analysis for Example (1). Alternatively, the verb phrase can be analyzed as a modifier of the initial noun phrase (RR clause; Example(4)) which is the correct analysis for Example (2). Further, Examples (1) and (2) are temporally ambiguous regarding the thematic relations of the sentential arguments. In Example (1), the initial noun phrase is the agent of the subsequent verb (main clause). In Example (2), the initial noun phrase is the patient of the verb of the RR clause (*sketched by the fencer*), and the agent of the main clause (*splashed the cellist*).

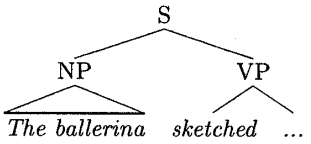


(3)



(4)

Provided factors such as plausibility, thematic fit, verb type (e.g., McRae, Spivey-Knowlton, & Tanenhaus, 1998), or frequency (e.g., Trueswell, 1996) do not bias toward one or the other construction during online comprehension, people initially prefer to adopt a main clause structure (see Example (5); Bever, 1970), and linguistic disambiguation only occurs later through the determiner or preposition of the second argument.



(5)

Whereas prior research has investigated the use of thematic fit (McRae et al., 1998), frequency (Trueswell, 1996), or referential context (e.g., Crain & Steedman, 1985) in the incremental structural disambiguation of the MV/RR ambiguity (for further reference see, e.g., Macdonald, 1994; Sedivy, 2002; Trueswell, Tanenhaus, & Garnsey, 1994), we investigate whether direct identification of depicted events as relevant through the utterance can enable a tightly time-locked structural disambiguation and thematic role assignment for this ambiguity.

In Experiment 1a, utterances such as Examples (1) and (2) were related to event depictions (Fig. 1). The main clause sentence in Example (1) described the ballerina-splashing-cellist event, whereas the RR clause in Example (2) described the fencer-sketching-ballerina event of Fig.1. If people can employ the verb for identification of the relevant depicted action during comprehension, as suggested by the findings for initially ambiguous German SVO/OVS sentences in Knoeferle et al. (2005), then we would expect the following for initially ambiguous English MV/RR sentences: For the main clause sentences (e.g., Example (1)), when the verb was *splashed,* and identified the ballerina-splashing action as relevant, we should find more anticipatory eye movements to the patient of the ballerina-splashing event (the cellist) than to the other agent (the fencer). For RR clause sentences when the verb was *sketched,* and identified the fencer-sketching action as the relevant action, we would expect to find more anticipatory eye movements to the agent of the sketching action (the fencer) than to the patient of the ballerina-splashing event (the cellist). Crucially, these eye movements should be anticipatory (Altmann & Kamide, 1999), and occur shortly after people encountered the verb and before disambiguation in the linguistic stimuli through the determiner or preposition of the second argument.

The MV/RR condition pair already described (e.g., Examples (1) and (2)) was one part of Experiment 1 that investigated verb-mediated identification of depicted events as relevant for structural disambiguation. Knoeferle et al. (2005) had further reported an early influence of depicted events on thematic role assignment when the events were mediated by subtle cues such as temporal adverbs rather than the verb. To test whether structural disambiguation through depicted events—when these are mediated by indirect linguistic cues—also takes

Fig. 1. Example image for Experiment 1.

place in close temporal coordination with encountering these cues, we designed a further condition pair. Recall that acquisition research suggests scene information is more relevant for the acquisition of concrete than abstract terms. In analogy, we proposed that scene information has a stronger influence on adult language comprehension when identified as relevant by concrete rather than abstract terms. If this were the case, then we would expect a weaker or less tightly coordinated utterance-mediated influence of depicted events for Experiment 1b. If the hypothesis is false, however, we would expect to see no difference between findings for Experiments 1a and 1b. Experiment 1b moreover enabled us to see whether findings for temporal adverbs from Knoeferle et al. generalize to another language, sentence type, and event cue (auxiliaries, see Examples (6) and (7)).

The ballerina will splash the cellist. (6)

The ballerina being sketched by the fencer splashed the cellist. (7)

We were specifically interested in whether preverbal auxiliaries *will* in Example (6), and *being* in Example (7) would establish a bias toward a future active or present tense passive event, respectively, before people encountered the verb. If the tense bias established by the auxiliaries is strong enough, a noun phrase–future auxiliary sequence (e.g., *The ballerina will*; Example (6)) should encourage participants to interpret the ballerina as the agent of a future action (splashing), and trigger anticipatory inspection of the patient of that action (the cellist; see Fig.1). Encountering, in contrast, a noun phrase followed by an auxiliary form that indicates the passive progressive aspect of an event (e.g., *The ballerina being*; Example (7)) should raise the expectation of a passive event, triggering anticipatory eye movements to the fencer as the agent of the event the princess undergoes. Based on findings from Experiment 3 in Knoeferle et al. (2005), we expected that these anticipatory eye movements to the patient and agent for *will* and *being* sentences, respectively, should occur shortly after the auxiliary has been encountered.

2.1. Method

2.1.1. Participants

Thirty-two participants from the University of Dundee (U.K.) were each paid 3 pounds for taking part in the experiment.

2.1.2. Materials

A set of 24 items was constructed. Fig. 2 shows a complete example image set for one item, and Table 1 displays the corresponding sentences. There were four conditions (Table 1: 1a, 1b, 2a, and 2b, for Fig. 2a). For counterbalancing reasons, a second image and an additional four sentences were added (Table 1: 1a´, 1b´, 2a´, and 2b´, and Fig. 2b), resulting in two images and eight sentences for one item.

We first discuss the four experimental conditions (Table 1: 1a, 1b, 2a, and 2b, and Fig. 2a), and subsequently explain the counterbalancing (Table 1: 1a´, 1b´, 2a´, and 2b´, and Fig. 2b). For the simple MV/RR conditions (1a and 1b), the ambiguous verb differed between the MV condition (e.g., *splashed*, Table 1, 1a) and the RR clause condition (e.g., *sketched,* Table 1, 1b).

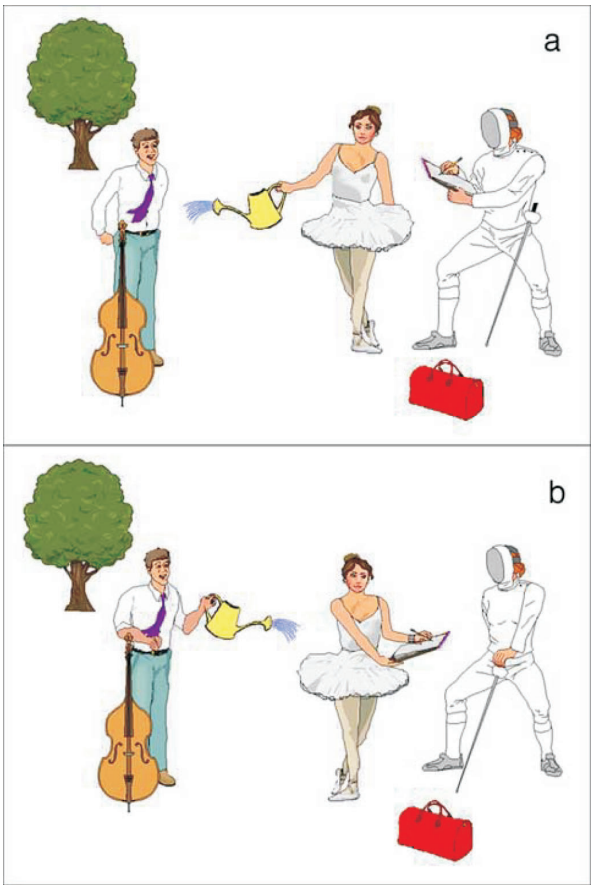


Fig. 2. Example image for sentences in Table 1 for Experiment 1.

The main clause sentences identified the ballerina-splashing-cellist event as relevant, and the RR clauses identified a different fencer-sketching-ballerina event as informative. Whereas for the main clause condition, the ambiguous character (the ballerina; see Fig. 2, Table 1, 1a) was the agent of a splashing action, she was the patient of the fencer-sketching event for the RR clause condition (Table 1, 1b). In contrast to McRae et al. (1998), stereotypicality or a plausibility bias were absent in our materials, and there was no frequency bias of the ambiguous verb that could be used for disambiguation (*t* test, $p > .3$; Trueswell, 1996). Thus, the only information type that was available for early thematic role assignment and structural disambiguation were depicted agent–action–patient events in the scene.

Images were kept constant for the second condition pair in which the verb was preceded by auxiliaries (*being/will* sentences; see Table 1, 2a and 2b). The auxiliaries were pretested for the strength of their bias by means of a sample from the British National Corpus. We randomly selected 250 occurrences each of both *will* and *being.* An analysis of their function in context revealed that *will* occurred in active sentences for 85.6% of the sample sentences, in passive constructions for 14%, and with other meanings in 0.4% of the samples. *Being* was followed by a

Table 1
Example sentence set for images in Fig. 2 in Experiment 1

Image	Condition	Sentences
Fig. 2a	MV	(1a) The ballerina splashed apparently the cellist in the white shirt.
Fig. 2a	RR	(1b) The ballerina sketched apparently by the fencer splashed the cellist.
Fig. 2a	*will*	(2a) The ballerina will apparently splash the cellist in the white shirt.
Fig. 2a	*being*	(2b) The ballerina being apparently sketched by the fencer splashed the cellist.
Fig. 2b	MV	(1a′) The ballerina sketched apparently the fencer in the white suit.
Fig. 2b	RR	(1b′) The ballerina splashed apparently by the cellist sketched the fencer.
Fig. 2b	*will*	(2a′) The ballerina will apparently sketch the fencer in the white suit.
Fig. 2b	*being*	(2b′) The ballerina being apparently splashed by the cellist sketched the fencer.

past participle in 78% of the occurrences, by an adjective in 20% of the cases, and had other meanings in 2% of the occurrences. Thus, *will* should bias toward an active sentence structure, and *being* toward a past participle construction.

To observe the influence of the verb-, or auxiliary-mediated depicted events in the eye movements after the verb or auxiliary had been encountered and prior to the onset of the disambiguating second argument, we introduced a filler adverb after the verb/auxiliary. Although this is a slightly unusual position for the adverb in those conditions where it followed the verb (1a and 1b; it typically precedes the verb), it is a constant across conditions and should therefore not affect the interpretation of relative eye movement proportions in the analysis.

For counterbalancing reasons, we added a second image version (Fig. 2b as a second version for Fig. 2a), and corresponding sentences (Table 1: 1a´, 1b´, 2a´, and 2b´) for each item image. This ensured that each target character (for Fig. 1 the cellist and the fencer) appeared once as patient and once as agent for each depicted action (and verb). Images were also included in their original and a mirrored version (the mirrored versions are not presented), resulting in 16 experimental lists. Each list was presented in two different randomizations so that all 32 participants saw the trials in an individually randomized order. As a result of the counterbalancing, conditions were matched for length and frequency of lemmas up to and including the second argument (t test, $p = 1$; Baayen, Pipenbrock, & Gulikers, 1995).

To exclude the influence of intonational cues on disambiguation, half of the item sentences were cross-spliced up to and including the adverb. For the simple MV/RR condition pair, the sentence beginning of 1a was spliced onto sentence 1b´ (Table 1). The sentence beginning of 1b was spliced on sentence 1a´. Sentence 1a´ was spliced in as the beginning of sentence 1b, and sentence 1b´ was spliced in as the beginning of sentence 1a. For the *will/being* condition pairs, we recorded an additional two sentences that had the same surface sequence of words as the experimental sentences, but had a different underlying structure. The sentence beginnings of these structurally different sentences were spliced in as beginnings for half of the *will* and *being* items to balance for intonational cues to the syntactic structure. For the *will* condition (Table 1, 2a), we recorded *The ballerina will apparently be splashed by the cellist*. The beginning of this passive sentence (*The ballerina will apparently*) was spliced in for sentences 2a and 2a´ (Table 1). For the *being* sentences of the example item (Table 1), we recorded the sentence *The ballerina being apparently very cheeky splashed the cellist*. The sentence beginning

of this utterance (up to and including the adverb) replaced the sentence beginnings of 2b and 2b´.

In addition to the 24 items, there were 32 filler trials, totaling 56 trials per participant. Each filler item consisted of a scene and a related utterance. Eight filler utterances started with an adverbial phrase and images showed two characters with only one being involved in an action; eight started with a noun phrase followed by *being* and an adjective, with images depicting four characters, two of which were involved in an action; eight had scenes that did not contain action depictions, with future tense sentences describing actions; eight had an initial noun phrase followed by a second coordinated noun phrase with images showing three characters, of which two were involved in an action. The fillers ensured that *will* and *being* did not always bias toward agent–patient or patient–agent events, respectively; that there was not always a noun phrase in the first position; that images did not always display a depicted action that could be used for comprehension, and that there were not always three characters in the scene. Experimental items were separated from one another by at least one filler item.

2.1.3. Procedure

An SMI EyeLink II head-mounted eye tracker monitored participants' eye movements with a sampling rate of 500 Hz. Images were presented on a 21-in. multiscan color monitor at a resolution of 1024 × 768 pixels concurrently with spoken sentences. Viewing was binocular and participants' head movements were unrestricted. We tracked only the dominant eye of each participant. For each trial, the image appeared 1,000 msec before utterance onset. Prior to the experiment, the experimenter instructed participants to listen to the sentences and to inspect the images, and to try to understand both sentences and depicted scenes. There was no other task. Participants were shown two example images and sentences. Next, the camera was set up and calibrated manually using a nine-point fixation stimulus. Calibration was repeated after approximately half of the trials. After calibration, the EyeLink software performed a validation; if validation was poor, the calibration procedure was repeated until validation was successful. The first experimental item for each participant was preceded by three filler items. Between the individual trials, a centrally located fixation dot appeared on the screen, which participants were asked to fixate. This allowed the eye tracking software to perform a drift correction if necessary. The entire experiment lasted approximately 30 min.

2.1.4. Analysis

The eye tracker software recorded the X-Y coordinates of participants' fixations. To analyze the output of the eye tracker, the visual scenes were coded into distinct regions on bitmap templates (1024 × 768 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 ballerina, the cellist, the fencer, and the distractor objects—a tree and bag—for Fig. 2). Characters were coded depending on their event role for the inferential and descriptive analyses (Figs. 4–6 and Figs. 8–10). For Fig. 2a, for instance, the ballerina would be coded as *ambiguous,* the fencer as *agent,* and the cellist as *patient.*

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 msec were pooled and incorporated into larger fixations; blinks and out-of-range fixations

were added to previous fixations. We report the mean proportions of fixations on scene entities over the course of the entire utterance (Figs. 3 and 7), as well as proportion of inspections and inferential analyses of the number of fixations for individual time regions (Figs. 4–6 and Figs. 8–10). We rely on these measures because previous studies on auditory sentence comprehension in visual environments have shown them to be closely linked to online comprehension processes (e.g., Sedivy et al., 1999; Tanenhaus et al., 1995).

For the time course presentation of the eye movement data (Figs. 3 and 7), the program computed the number of inspections that fell within a given time slot for each object. For example, if an inspection on an object started at 1,000 msec and lasted until 1,625 msec, and time slots were 250 msec, then the program scored one inspection on that object for the 1,000–1,250 msec time slot, one inspection for the 1,250–1,500 msec time slot, and finally one inspection for the 1,500–1,750 msec time slot (i.e., the end of the inspection fell still within the 1,500–1,750 msec slot). For the subsequent slot from 1,750 to 2,000 msec the program would score zero inspections unless a new inspection started within that slot. Figs. 3 and 7 plot the mean of the proportion of inspections per time slot, separately for each sentence condition and role of character. The word onsets marked on the graphs represent the average of word onsets for the individual item trials (see average durations given later).

The data presented for the individual time regions (Figs. 4–6 and Figs. 8–10) are, in contrast, based on exact computations of these regions for each individual trial. Word onsets had been marked for the first noun phrase, the verb or auxiliary, the adverb, and the second argument in each item speech file. We computed the proportion of cases per sentence condition for which inspections started within a time region. This calculation differs from the computation for the time curve presentation, where an inspection was counted when it fell within a given time slot. Owing to this difference in computation, there may be slight differences between the two types of data presentation in some cases. When an inspection started in one slot and lasted into the following time slot, it would be counted as an inspection during both periods in the time curve presentation (Figs. 3 and 7). For the graphs of the individual utterance regions (Figs. 4–6 and Figs. 8–10), it would only be counted for the region in which it started.

For the inferential analysis of inspection proportions during a time region we used hierarchical log-linear models. These combine characteristics of a standard cross-tabulation chi-square test with those of analysis of variance (ANOVA). Log-linear models neither rely on 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). Inspection counts for a time region were adjusted to factor combinations of target character (patient, agent), sentence condition (MV/RR for conditions 1a and 1b, and *will*/*being* for conditions 2a and 2b of Table 1) and either participants ($N = 32$) 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).

For the simple MV/RR conditions the analysis time regions were the verb, the adverb, and the second argument. Because we expected no effect prior to the verb, we do not present inferential analyses for the preverbal noun phrase region. The time curves, however, show the eye movements over the entire course of the utterance. The VERB region extends from 180 msec after verb onset to the onset of the adverb ($M = 486$ msec for MV and $M = 485$ msec for the RR condition). This choice of region boundaries should maximize the possibility of finding an

even earlier (and thus more closely verb-mediated) effect of the depicted events than the postverbal effect reported in Knoeferle et al. (2005). We chose the onset of the region 180 msec into the verb because this corresponds to the earliest point at which eye movements could reflect the verb-mediated effects of depicted events (e.g., Altmann & Kamide, 2004; Matin, Shao, & Boff, 1993), and at the same time effects of the first noun phrase are starting to decay. While exploring this possibility, we chose the postverbal ADV as the main analysis region (see Knoeferle et al., 2005). This region stretched from the onset of the adverb to the onset of the second argument (M = 794 msec for both MV and RR conditions). The SECOND ARGUMENT region represents the second noun phrase for MV, and the prepositional phrase for RR clauses. The mean duration for the second argument in main verb sentences was 690 msec, and this was the region used for analyses. For the prepositional phrase in RR clauses the mean duration was 887 msec. Because the second argument region for the RR condition greatly exceeded the length of the second argument region for the MV condition, a difference in the proportions of inspections to the agent in the RR condition as compared to the MV condition could be explained in terms of the longer analysis region for the RR condition. We adjusted for the inequality in length between the two conditions by subtracting the length difference (197 msec) from the offset of the prepositional phrase for each trial in the RR condition. The mean duration of the prepositional phrase after this subtraction was 690 msec, and this was the region used for analyses in the RR condition (SECOND ARGUMENT).

For the *will/being* conditions, we chose the auxiliary and adverb, the verb, and the second argument as analysis regions. As the auxiliaries were very short, and the earliest effect could occur on the postauxiliary adverb region, we collapsed the auxiliary and adverb into one region. Following the same reasoning as for the VERB region, the AUXADV region starts 180 msec after the onset of *will/being* and lasts until the onset of the verb in the ambiguous clause (874 msec for *will* and 845 msec for *being* sentences). The verb region (VERB) for the *will/being* conditions was positioned after the AUXADV region and stretched from the onset of the verb to the onset of the second argument (582 msec for *will* and 622 msec for the *being* condition). In this case (unlike for the verb in the MV/RR conditions), we chose the entire duration of the verb as the region, as the preceding auxiliary and adverb should already have triggered disambiguation for the *will/being* conditions; thus maximizing chances of finding evidence for structural disambiguation was no longer an issue. The SECOND ARGUMENT region comprises the second noun phrase for the *will* and the prepositional phrase for the *being* condition. Just as for the MV/RR condition pair, the prepositional phrase exceeded the noun phrase in length. To adjust for this difference, the length difference between the noun phrase and the prepositional phrase (211 msec) was deducted from the offset of the prepositional phrase for all trials. The resulting prepositional phrase region was 674 msec and matched the duration of the noun phrase (M = 674 msec).

2.2. Results and discussion

Figures 3 and 7 present an overview of the time course of eye movements during presentation of the utterance. Fig. 3a displays inspection proportions to the agent and patient, and Fig. 3b to the ambiguous character for the simple MV/RR conditions (Experiment 1a, Table 1, 1a and 1b). Fig. 7a and Fig. 7b show eye movements to the agent/patient and to the ambiguous character, respectively, for the *will/being* conditions (Experiment 1b, Table 1, 2a and 2b).

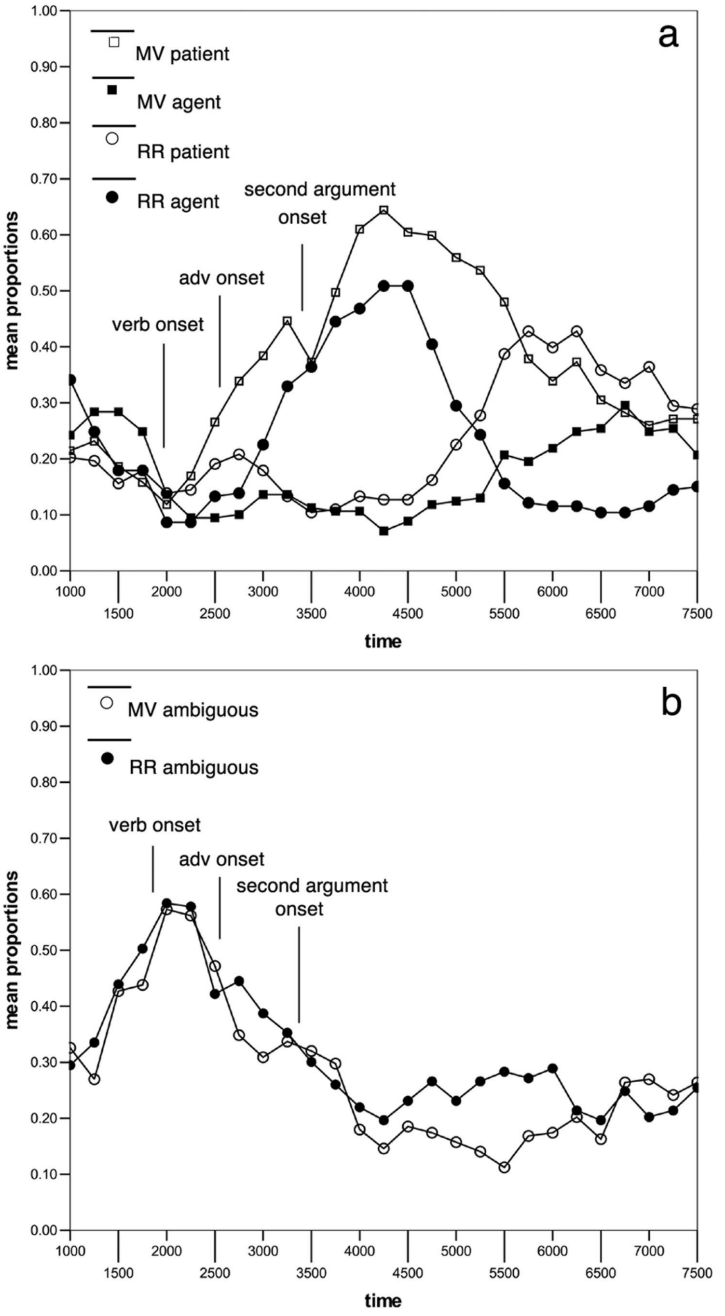


Fig. 3. Time course of the eye movement data for Experiment 1 (MV/RR condition) showing the mean proportion of inspections to entities from the onset of the spoken stimuli in time frames of 250 msec.

Inspections to the background and the distractor objects were omitted from the time course presentation. All of these are, however, included in the graphs that show the proportions of inspections during the individual analysis regions (Figs. 4–6 and Figs. 8–10). In the discussion we only comment on the eye movement pattern up to and including the second argument because the remainder of the sentence was not of interest for the research questions we posed, and was not matched for length and frequency.

2.2.1. Experiment 1a: Simple MV/RR clause condition

Prior to verb onset, when people are listening to the first noun phrase, and during the early verb time region, most inspections go to the ambiguous character (the ballerina), which is the referent of the first noun phrase (see Fig. 3b). As people hear the verb and the subsequent adverb, eye movements to the ambiguous character start to decline. In the eye-gaze pattern to the ambiguous character during the later utterance (from 4,500 msec), we observe a higher proportion of inspections to the ambiguous character for the RR in comparison with the MV condition. Because item sentences were, however, only matched for length and frequency up to and including the second argument, we cannot further interpret this pattern.

We now focus our attention on Fig. 3a, which displays the gaze pattern to the patient (the cellist) and to the agent (the fencer; see Fig. 1a and Table 1, 1a and 1b). The graph shows that there are more inspections to the patient than agent for both the MV and RR conditions during the verb. This replicates findings from Knoeferle et al. (2005), and could be due to visual factors (the first-named ambiguous character is oriented toward the patient), or linguistic expectation of the patient triggered by a main clause preference (e.g., Bever, 1970).

Between the onset of the adverb and the onset of the second argument, fixation patterns for the MV condition in comparison with the RR condition diverge. When the verb described the ballerina-splashing action, participants looked more at the patient of that action (the cellist) than at the agent of the sketching event (the fencer) shortly after verb offset. In contrast, when hearing a verb that identified the action performed by the other agent on the ambiguous character as relevant (fencer-sketching-ballerina) in the RR condition, participants started looking more at the fencer than at the patient of the ballerina-splashing action (the cellist). At the same time their inspection of the cellist (patient) decreased rapidly for the RR condition.

During the second argument region the clear disambiguation pattern observed during the postverbal adverb region continues. The important conclusion from the discussion of Fig. 3a is that the pattern of anticipatory eye movements to the patient (for MV) and to the agent (for RR) suggests clear disambiguation early during the postverbal adverb region and thus closely coordinated with the identification of the depicted events as relevant through the verb.

This key finding was confirmed by the inferential analyses. Whereas the time curves present the eye movement pattern over the course of the entire utterance, the inferential analyses focus only on the regions with the theoretically most relevant effects.

During the VERB region, there was a main effect of more inspections to the patient than agent for both sentence conditions (MV, RR), $LR\chi^2(\text{subj}) = 11.45$, $df = 1$, $p < .001$; $p > .2$ by items, in the absence of an interaction ($ps > 0.1$; see Fig. 4).

During the ADV region (see Fig. 5), log-linear analyses revealed a significant interaction between target character (patient, agent) and sentence condition (MV, RR), $LR\chi^2(\text{subj}) = 60.02$, $df = 1$, $p < .0001$. Although the analysis by items was not reliable for the ADV region, analyses for a

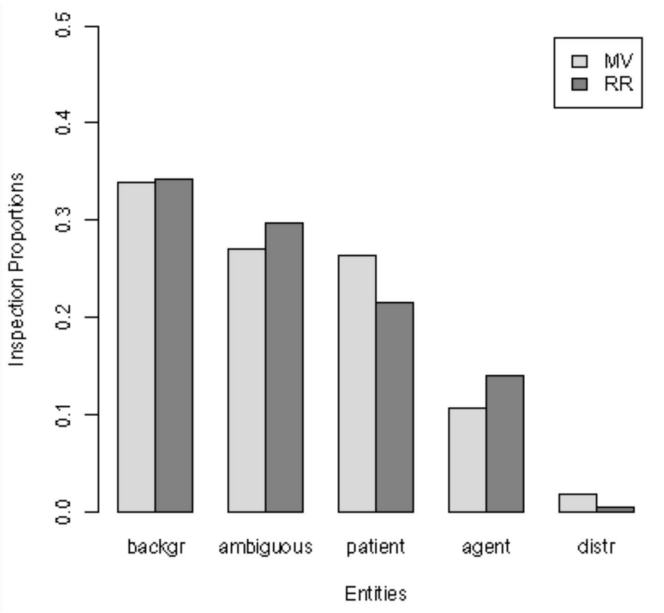


Fig. 4. Percentage of inspections to entities during the VERB region for the MV/RR condition in Experiment 1a.

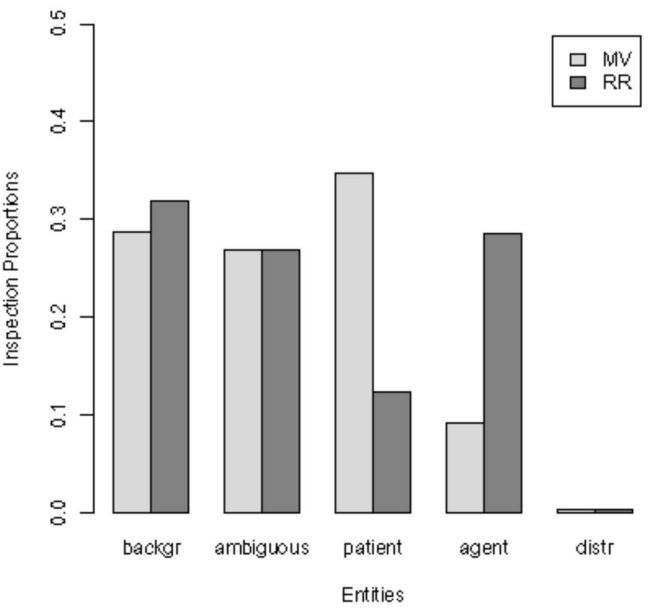


Fig. 5. Percentage of inspections to entities during the ADV region for the MV/RR condition in Experiment 1a.

shorter region that stretched from the onset of the adverb to its offset for each trial revealed a significant interaction of target character (patient, agent) and sentence condition (MV, RR), $LR\chi^2$(subj) = 17.15, df = 1, $p < .0001$; $LR\chi^2$(item) = 47.69, df = 1, $p < .0001$. The interaction was due to a higher proportion of inspections to the patient for the MV in comparison with the RR condition, $LR\chi^2$(subj) = 40.40, df = 1, $p < .0001$; $LR\chi^2$(item) = 42.50, df = 1, $p < .0001$, and to a greater percentage of inspection to the agent in the RR compared with the MV condition, $LR\chi^2$(subj) = 35.52, df = 1, $p < .0001$; $LR\chi^2$(item) = 34.00, df = 1, $p < .0001$.

For the SECOND ARGUMENT region (see Fig. 6), we found a significant interaction between target character (agent, patient), and sentence condition (MV, RR), $LR\chi^2$(subj) = 60.88, df = 1, $p < .0001$; $LR\chi^2$(item) = 59.35, df = 1, $p < .0001$. Contrasts showed it was due to a significantly higher proportion of patient inspections for MV in comparison with RR sentences, $LR\chi^2$(subj) = 36.63, df = 1, $p < .0001$; $LR\chi^2$(item) = 36.74, df = 1, $p < .0001$, and due to a higher percentage of inspections on the agent for RR compared to MV sentences, $LR\chi^2$(subj) = 39.03, df = 1, $p < .0001$; $LR\chi^2$(item) = 34.79, df = 1, $p < .0001$.

In sum, our findings from the MV/RR ambiguity (Experiment 1a) provide strong evidence for the tight temporal coordination between verb-mediation of depicted events and the influence of those events on comprehension. Early structural disambiguation for the MV/RR sentence occurred shortly after people had heard the verb and prior to disambiguation through the determiner or preposition on the second argument.

2.2.2. *Experiment 1b: Will/being conditions*

Prior to the onset of the verb, most people look at the ambiguous character (see Fig. 7b). There are more inspections to the ambiguous character in the *being* condition than in the *will*

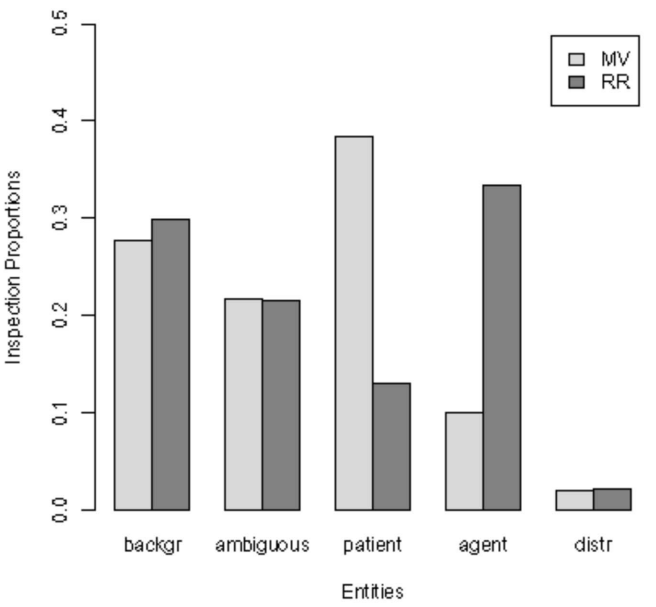


Fig. 6. Percentage of inspections to entities during the SECOND ARGUMENT region for the MV/RR condition in Experiment 1a.

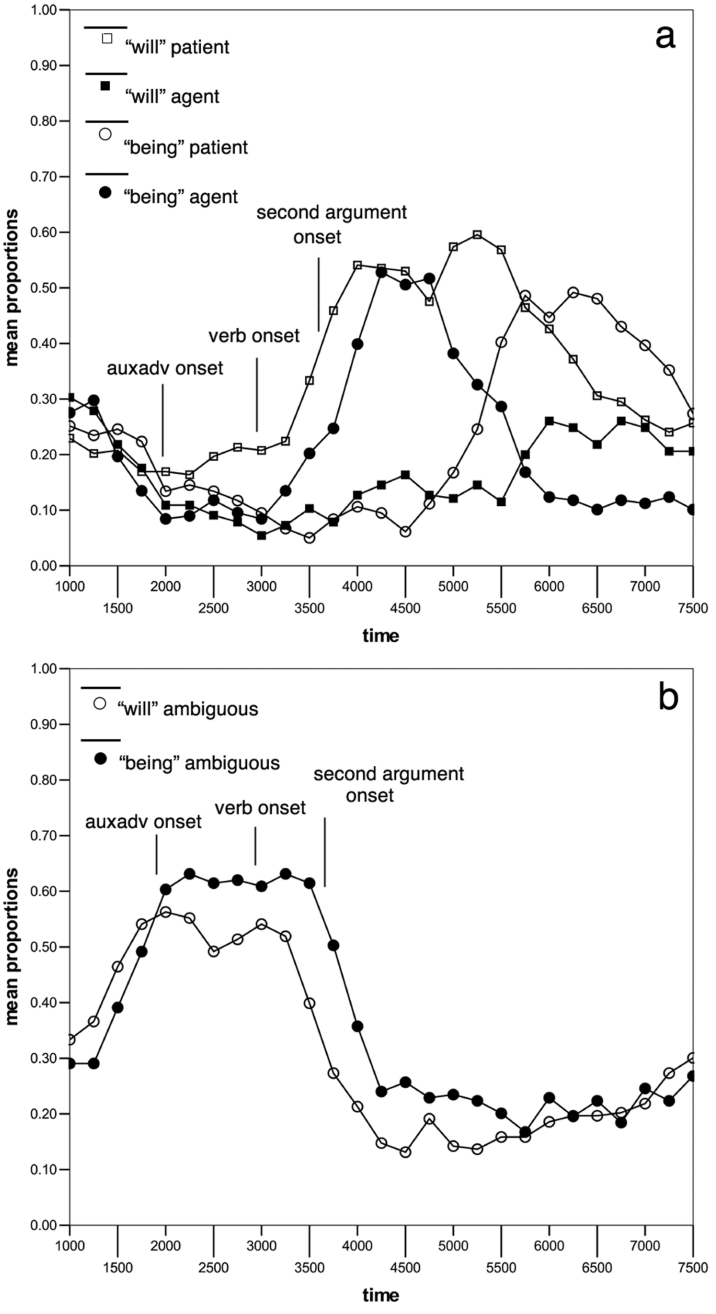


Fig. 7. Time course of the eye movement data for Experiment 1 (*will/being* condition) showing the mean proportion of inspections to entities from the onset of the spoken stimuli in time frames of 250 msec.

condition, starting with the onset of the auxiliary and continuing throughout the verb and second argument. This descriptive pattern was, however, not significant in the inferential analyses for the individual analysis regions.

Fig. 7a shows that after the onset of the auxiliary, looks to the patient increase for the *will* condition. This could be due to a general structural preference for an active sentence structure, the specific bias of the future auxiliary, or both (recall the MV condition of Table 1, 1a). Crucially, in the *will/being* comparison, such a patient preference occurs only for the active sentence condition, and is absent for the *being* condition. This descriptive finding contrasts with results from the experiments presented by Knoeferle et al. (2005), and also with findings from the MV/RR comparison (Experiment 1a). In all of these experiments, shortly after the offset of the first noun phrase, inspection of the patient increased always for both the favored and disfavored sentence type. The absence of such an increase when people heard *being* suggests that *being* was interpreted immediately as biasing toward a passive structure, and suppressed early looks to the patient. This pattern was, however, not confirmed by an interaction in the inferential analyses (see later). Rather, the inferential analyses reveal a main effect of target character just as in Experiment 1a. There was further no strong increase of looks to the agent for the *being* condition in comparison with the *will* condition prior to the onset of the verb (see Fig. 7a).

We find evidence for disambiguation while the verb is encountered. With verb onset, looks to the agent for the *being* condition exceed looks to the patient for that condition. For the *will* condition, the increased proportion of inspections to the patient in comparison with inspection of the agent condition continues. This gaze pattern during the verb can be interpreted in two ways. Either preverbal auxiliary cues alone made the relevant depicted event available, or the auxiliary cues activated event structures, but required the additional influence of the verb to trigger clear disambiguation. The view that the auxiliaries alone activated event structures is supported by the observation that disambiguation through verb-mediated events had only occurred after the verb region for the MV/RR conditions in Experiment 1a (see Fig. 3a) as well as for Experiments 1 and 2 in Knoeferle et al. (2005). The second interpretation is supported by findings from Kamide, Scheepers, and Altmann (2003), who reported the combined effects of tense cues (*will/will be*), verb information, and world knowledge during comprehension of sentences such as *The hare will eat/will be eaten the cabbage/by the fox*. Further research is required to tease apart these alternative interpretations.

We now turn to the inferential analysis. The main finding—the early disambiguation observed in the gaze pattern during the lifetime of the verb—was crucially confirmed in the inferential analyses. For the AUXADV region, log-linear analyses revealed a main effect of target character (patient, agent), $LR\chi^2(\text{subj}) = 4.73$, $df = 1$, $p < .05$; $p > .2$ by items, in the absence of an interaction (ps > .2; Fig. 8). The difference in the inspections to the ambiguous character for MV compared with RR clause sentences that appeared in the time curves (see Fig. 7b) was not confirmed by the analyses for the AUXADV region (ps > .5).

During the VERB region (Fig. 9), log-linear analyses showed a main effect of target character (patient, agent) with more inspections to the patient than to the agent, $LR\chi^2(\text{subj}) = 11.50$, $df = 1$, $p < .001$; $LR\chi^2(\text{item}) = 11.50$, $df = 1$, $p < .001$. Analyses further revealed a significant interaction between target character (patient, agent) and sentence condition (*will, being*), $LR\chi^2(\text{subj}) = 30.16$, $df = 1$, $p < .0001$; $LR\chi^2(\text{item}) = 22.38$, $df = 1$, $p < .0001$. Contrasts showed that this resulted from a higher proportion of inspections to the patient for the *will* as opposed

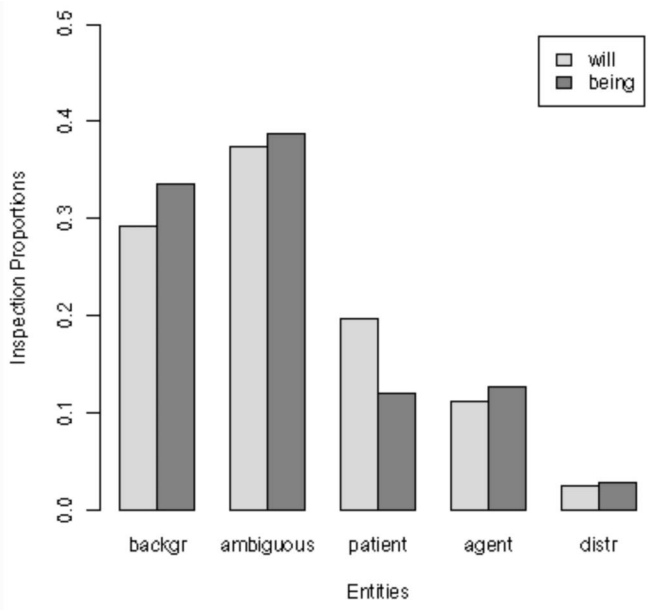


Fig. 8. Percentage of inspections to entities during the AUXADV region for the *will/being* condition in Experiment 1b.

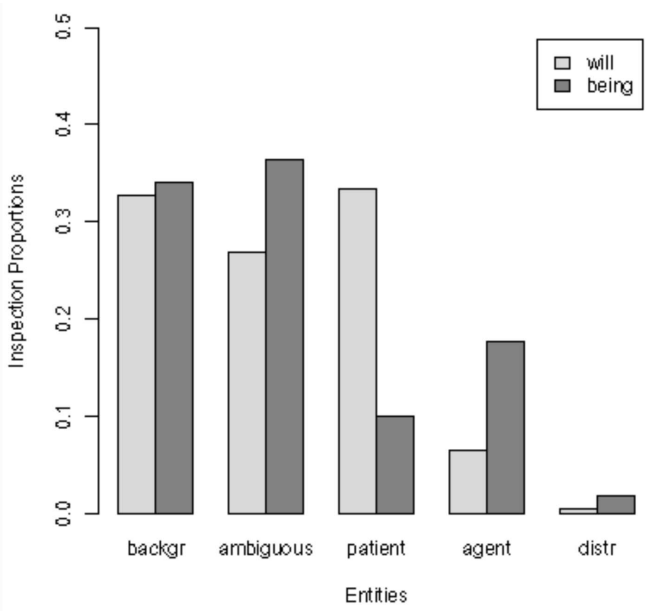


Fig. 9. Percentage of inspections to entities during the VERB region for the *will/being* condition in Experiment 1b.

to the *being* condition, $LR\chi^2$(subj) = 34.59, $df = 1$, $p < .0001$; $LR\chi^2$(item) = 30.37, $df = 1$, $p < .0001$. The increased proportion of inspections to the agent for the *being* in comparison with the *will* condition were not significant (*p*s > .2). The descriptively higher proportion of inspections to the ambiguous character for the *being* in comparison with *will* sentences was not significant (*p*s > .8).

For the SECOND ARGUMENT region (Fig. 10), there was a significant interaction of target character (patient, agent) and sentence condition (*will, being*), $LR\chi^2$(subj) = 79.20, $df = 1$, $p < .0001$; $LR\chi^2$(item) = 90.96, $df = 1$, $p < .0001$. The interaction resulted from a higher proportion of inspections to the patient for the *will* in comparison with the *being* condition, $LR\chi^2$(subj) = 50.38, $df = 1$, $p < .0001$; $LR\chi^2$(item) = 54.72, $df = 1$, $p < .0001$, and from a higher percentage of inspections to the agent for the *being* as opposed to the *will* condition, $LR\chi^2$(subj) = 59.81, $df = 1$, $p < .0001$; $LR\chi^2$(item) = 63.67, $df = 1$, $p < 0.0001$.

Findings from Experiment 1b suggest that indirect auxiliary tense cues facilitate the verb-mediated use of scene events for rapid disambiguation. Disambiguation occurred one region earlier (during the lifetime of the verb) than when no tense cues preceded the verb. The pattern for disambiguation that we observed during the lexical verb was further confirmed during disambiguation due to the determiner or preposition on the second argument.

The main insight from Experiment 1 is the strong evidence for the claim that identification of depicted agent–action–patient events as relevant for comprehension through verb reference is closely temporally coordinated with the influence of these events on incremental thematic role assignment and structural disambiguation. Findings for the indirect tense cues—although suggestive of a tight temporal coordination—require further research to clarify the extent to which a tight coordination holds when depicted events are identified as relevant by indirect cues.

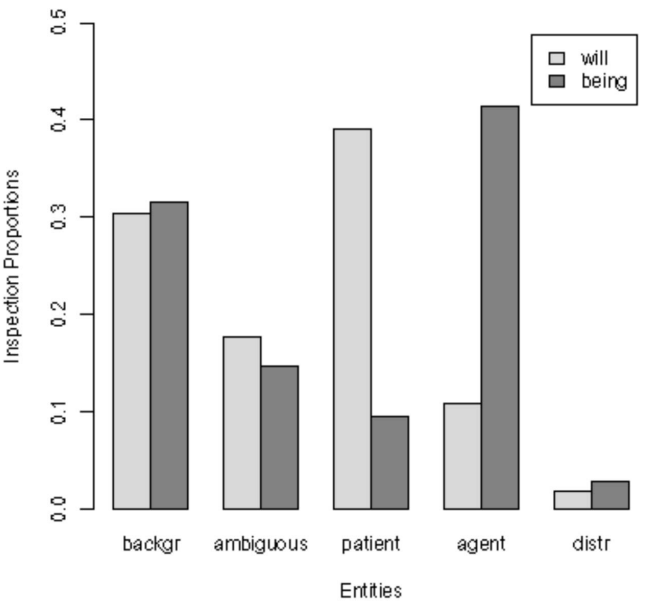


Fig. 10. Percentage of inspections to entities for the SECOND ARGUMENT region for the *will/being* condition in Experiment 1b.

3. Experiment 2

In Experiment 2 we further investigate the verb-mediated influence of depicted events. Although allowing us to confirm findings on the tight temporal coordination (Experiment 1), the focus of Experiment 2 was on examining the relative importance of depicted events in comparison with stereotypical thematic role knowledge. Our first aim was to replicate that both stored world knowledge about likely thematic role fillers (Kamide, Scheepers, & Altmann, 2003) and depicted events (Experiment 1, and Knoeferle et al., 2005) can influence incremental thematic role assignment when only one of them is identified as relevant by the verb in the utterance. Second, within the same experiment we explored the relative importance of depicted events in comparison with world knowledge about likely role fillers for incremental thematic role assignment. To test this, utterances identified both a stereotypical agent and a different agent of a depicted action as relevant for comprehension. Crucially, because the utterance in this case did not uniquely identify only one scene entity as relevant for utterance interpretation, the comprehension system was forced to choose between the two available and relevant agents for thematic interpretation of the utterance.

We exploited scenes such as the one presented in Fig. 11. A scene shows three characters (e.g., a pilot, a wizard, and a detective). The middle character is always a patient (i.e., not performing an action), and the other two characters (the wizard and the detective) always have an agent role. Each of these two characters (the wizard and the detective) can be qualified as agent in two respects: through stereotypical knowledge about what wizards and detectives typically do, and through the immediately depicted event that they are performing. The wizard is a stereotypical agent of a jinxing action, and depicted as performing a spying action. The detective is a stereotypical agent of a spying action, and is depicted as serving food to the pilot. Thus, stereotypical knowledge (e.g., wizard-jinxing) and depicted events (e.g., detective-serving-food) uniquely identify a (different) agent on an image as relevant for comprehension. In addition, one agent (e.g., the wizard) is depicted as performing an action (spying) that is a stereotypical action of the other agent (the detective; see Fig. 11).

Item sentences in Experiment 2 were unambiguous OVS sentences that related to patient–action–agent events (Fig. 11). Recall that German has a rich case-marking system where

Fig. 11. Example image for Experiment 2.

grammatical function is usually indicated by unambiguous case morphemes. Word order constraints are less rigid in German than in English, and both SVO and OVS order are grammatical, with SVO being the preferred reading (e.g., Hemforth, 1993).

For the image in Fig. 11, an OVS sentence fragment such as *Den Piloten (ACC) verzaubert* … ('The pilot (ACC) jinxes … ') identifies one available scene entity as relevant. Accusative case-marking on the determiner of the first noun phrase permits assignment of a patient role to that noun phrase while establishing reference to the pilot in the scene. On hearing the verb, *verzaubert* ('jinxes'), our knowledge that a wizard is a likely jinxing agent can combine with the fact that the wizard is the only stereotypical agent for a jinxing action. Findings by Kamide, Scheepers, and Altmann (2003) suggest that the combination of case-marking, verb-based thematic role knowledge, and available role fillers allows us to anticipate the wizard as a likely-to-be-mentioned agent (Fig. 11). Based on their findings, we expected anticipatory eye movements to the wizard shortly after people heard the verb.

In contrast, when people hear *Den Piloten (ACC) verköstigt …* ('The pilot (ACC) serves food to … '), verb-based thematic role knowledge of stereotypical agents of a serving-food action (e.g., a cook) cannot enable incremental thematic role assignment, because the scene contains no such entity. However, the scene does contain a depicted food-serving event performed by the detective that can enable thematic interpretation of the utterance. Based on findings from Experiment 1 of this article, and from Knoeferle et al. (2005), we would expect anticipatory eye movements to the agent of the verb-mediated depicted action (serving food) once people have heard the verb *verköstigt* ('serve-food-to'; Fig. 11). It should be noted that in both of these examples the verb uniquely identified either a depicted action and its (nontypical) agent as relevant, or it guided attention toward a stereotypical agent of the verb that performed an unrelated action.

Imagine we heard instead *Den Piloten (ACC) bespitzelt* ('The pilot (ACC) spies-on') In this case, the verb does not uniquely identify either a depicted action and its (nontypical) agent or a stereotypical (nondepicted) action and its agent as relevant. Rather, it identifies both a stereotypical agent (the detective), and an immediately depicted agent of a spying event (the wizard; Fig. 11) as likely target agents. Do listeners rely more on extracting thematic role relations from stereotypical knowledge provided by the utterance ('spy-on' + world knowledge → detective), or do they rely on role relations proffered by the scene ('spy-on' + wizard-spying event → wizard) in incremental interpretation? For the ambiguous *bespitzeln* example ('spy-on'), information about who does what to whom in the depicted events conflicts with stereotypical knowledge of who does what to whom. The comprehension system has to choose between two available yet conflicting types of information in determining the appropriate agent in the scene.

After people heard *Den Piloten (ACC) bespitzelt* ('The pilot (ACC) spies on'), no early interaction is expected in the gaze pattern because the utterance beginning is identical for these two conditions, and only the second noun phrase either identifies the agent of a stereotypical or depicted action as relevant (e.g., *bespitzelt der Detektiv/Zauberer (NOM)*, 'spies on the detective/wizard (NOM)'). Rather, we expect an early main effect with the interesting point being its directionality: If the comprehension system relies on stored thematic role knowledge in preference to depicted events we should find a higher percentage of anticipatory inspections to the stereotypical spying agent (the detective) than to the depicted spying agent shortly after the verb (Fig. 11). Alternatively, if the comprehension system prefers to rely on role relations proffered by the immediate scene in incremental thematic interpretation, we expect a higher proportion of anticipatory eye movements to the depicted than to the stereotypical agent after the

verb *bespitzeln* 'spy on' has been encountered. A third option is that we find no clear preference for either of these information types, but rather that competition between these two information types in identifying the correct agent as relevant creates comprehension difficulty. In this case, we expect an even distribution of inspections between the detective (the stereotypical agent of a spying event) and the wizard (depicted as the agent of a spying event). When the second noun phrase is encountered, and directly identifies which agent is relevant, we expect an interaction, resulting from the same pattern of eye movements as for the sentences where the verb uniquely identifies only one agent in the scene as informative for comprehension.

3.1. Method

3.1.1. Participants

Twenty-four German native speakers with normal or corrected-to-normal vision each received 5 euros for participation.

3.1.2. Materials, design, and procedure

There were 24 items and four experimental conditions (Table 2: a1, a2, b1, b2). For counterbalancing reasons, however, one item comprised eight sentences and two images. We thus created 48 images using commercially available clip art and graphic software. Further counterbal-

Table 2
Example item sentence set for Fig. 12

Image	Condition		Sentences
Fig. 12a	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. 12a	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. 12a	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. 12a	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. 12b	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. 12b	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. 12b	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. 12b	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.

ancing was performed between two items (see Table 2 and Fig. 12; Table 3 and Fig. 13). We first describe the four experimental conditions for one image, and then proceed to explain the counterbalancing.

An example image showed two agents (e.g., a wizard, and a detective), each performing an action on a patient (e.g., the pilot; see Fig. 12a). The depicted events provided information about role relations (e.g., wizard-spying-on-pilot and detective-serving-food-to-pilot, see Fig. 12a). In addition, each agent provided stereotypical thematic role knowledge (e.g., a detective is a stereotypical agent of a spying action, and a wizard a stereotypical agent of a jinxing action). We refer to the agent of a depicted event as the *depicted agent* and to a character that is identified as a relevant agent through verb-based stored thematic knowledge as the *stereotypical agent.*

All sentences were unambiguous OVS sentences. They always started with an unambiguously accusative case-marked noun phrase referring to a patient role filler (Fig. 12a, the pilot). Although the OVS structure is noncanonical in German, this was a constant across conditions.

Manipulation of the verb created four conditions, crossing the factors target type (depicted, stereotypical) with identification (unique, ambiguous). *Identification* refers to whether the

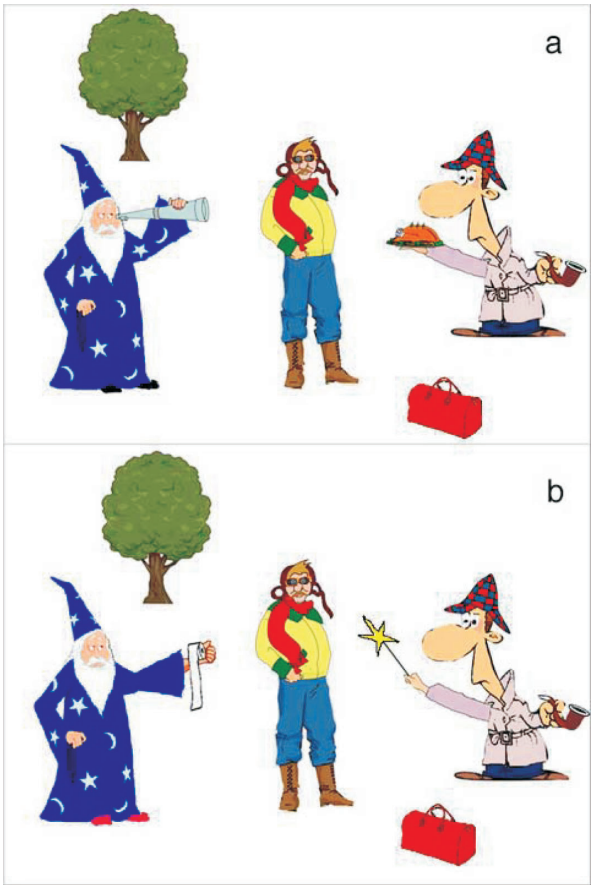


Fig. 12. Example image for sentences in Table 2 for Experiment 2.

Table 3
Example item sentence set for Fig. 13

Image	Condition		Sentences
Fig. 13a	Unique identification & depicted target	(a1)	Den Touristen verzaubert gleich der Medikus. The tourist (PAT.) jinxes soon the doctor. The doctor will soon jinx the tourist.
Fig. 13a	Unique identification & stereotypical target	(a2)	Den Touristen verköstigt gleich der Kneipenwirt. The tourist (PAT.) serves-food-to soon the innkeeper. The innkeeper will soon serve food to the tourist.
Fig. 13a	Ambiguous identification & depicted target	(b1)	Den Touristen bandgagiert gleich der Kneipenwirt. The tourist (PAT.) bandages soon the innkeeper. The innkeeper will soon bandage the tourist.
Fig. 13a	Ambiguous identification & stereotypical target	(b2)	Den Touristen bandgagiert gleich der Medikus. The tourist (PAT.) bandages soon the doctor. The doctor will soon bandage the tourist.
Fig. 13b	Unique identification & depicted target	(a1')	Den Touristen bespitzelt gleich der Kneipenwirt. The tourist (PAT.) spies-on soon the innkeeper. The innkeeper will soon spy on the tourist.
Fig. 13b	Unique identification & stereotypical target	(a2')	Den Touristen bandagiert gleich der Medikus. The tourist (PAT.) bandages soon the doctor. The doctor will soon bandage the tourist.
Fig. 13b	Ambiguous identification & depicted target	(b1')	Den Touristen verköstigt gleich der Medikus. The tourist (PAT.) serves-food-to soon the doctor. The doctor will soon serve food to the tourist.
Fig. 13b	Ambiguous identification & stereotypical target	(b2')	Den Touristen verköstigt gleich der Kneipenwirt. The tourist (PAT.) serves-food-to soon the innkeeper. The innkeeper will soon serve food to the tourist.

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 type that is identified as relevant by the second noun phrase. When referring to the target agent identified by the second noun phrase we will use target type. When we refer, in contrast, to the scene entities, we will use the expression target agent (depicted agent and stereotypical agent).

We designed the materials so that plausibility biases introduced by the verb or noun phrases, depiction biases introduced by scene characters, or position biases (whether an agent was to the left or right of the patient) were counterbalanced. Two image versions were created for each item (e.g., Fig. 12a and 12b), and we recorded four sentences for each image version (e.g., for Fig. 12a, the sentences a1, a2, b1, and b2 in Table 2). The two image versions (henceforth image direction) ensured that whether a target agent was to the left or right of the patient on an im-

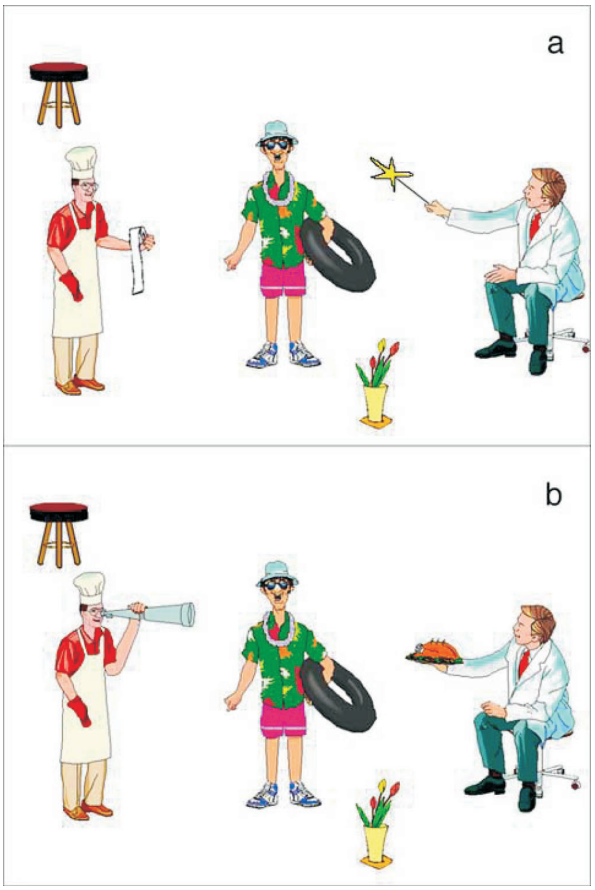


Fig. 13. Example image for sentences in Table 3 for Experiment 2.

age was balanced. For the unique identification conditions and Fig. 12a (Direction 1), for example, the stereotypical agent (jinx, wizard) was left of the patient, and the depicted agent (detective-serving-food) was to the pilot's right. For the unique conditions and Fig. 12b (Direction 2), the stereotypical agent (detective) was to the right on the image, and the depicted agent (wizard-bandaging) was leftmost. For the ambiguous identification conditions, the agent identified as target by stereotypical knowledge ('spy-on,' detective) was rightmost, and the agent depicted as performing a spying action was leftmost (wizard-spying; Fig. 12a). For Fig. 12b this was reversed.

The balancing further ensured that each verb identified once the agent of a depicted event as relevant (e.g., *verzaubern*, 'jinx,' detective-jinxing; Table 2, b1´, and Fig. 12b), and once a stereotypical agent ('jinx,' wizard-jinxing; Table 2 a2, and Fig. 12a). It further ensured that each of the target agents (wizard, detective) was once identified as a potential agent through verb-mediated depicted actions (e.g., wizard-bandaging; Table 2, a1´, and Fig. 12b) and through stereotypical verb-based knowledge (e.g., a wizard is a stereotypical agent of a jinxing action, Table 2, a2, and Fig. 12a). Each agent was once identified uniquely as relevant through

the utterance (e.g., the detective in a1), and once identification of the relevant agent was ambiguous (e.g., the detective in b2). Finally, each of the four verbs in an item set was used in every experiment condition between two item sets (e.g., *verköstigt*, 'serves-food-to'; see Table 2, a1, and Table 3, a2, b1´, and b2´). As a result of the counterbalancing, conditions were matched for length (number of spoken syllables), and frequency of lemmas within an item (*t* test, $p = 1$; Baayen et al., 1995).

To ensure further that the plausibility manipulations were strong enough, and that the character that was the stereotypical agent for a verb (detective, *spy-on*) was a more plausible agent for that verb than the other character (wizard; Fig. 12a), we carried out an offline plausibility rating test. For both target agents of an image (the detective and the wizard; Fig. 12a), participants would see three unambiguous SVO sentences such as "The … (subj) serves-food-to/spies-on/jinxes the pilot (obj)." There were six conditions depending on whether the rated character was a stereotypical agent or not (stereotypical agent, Table 4); whether the rated character was depicted as performing the action identified as relevant by the verb (depicted-action agent, Table 4); and whether the other character was a competitor (i.e., a plausible agent or the agent of a depicted action for that verb) or not (competitor, Table 4). The design ensured that participants provided ratings for all the verb–target agent combinations that appeared in Experiment 2. To give an example for the preceding three sentences and Fig. 12a: The wizard was rated for sentences with the verbs "serves food to", "jinxes", and "spies on" (Conditions 1, 4, and 5 in Table 4, respectively); the detective was rated for sentences with the same verbs jinxes, spies on, and serves food to (in Conditions 2, 3, and 6, in Table 4). Participants were asked to rate each of the two agents on an image on a scale from 1 (*not plausible*) to 7 (*highly plausible*) with respect to how plausible it was for that character to perform the action indicated by the verb on the patient.

Results showed that our plausibility manipulation was strong (see Table 4). Only when the character was a stereotypical target (e.g., wizard for "jinx" and detective for "spy on"; Fig. 12a, Table 2, a2, b2) were mean plausibility ratings high. This was the case both when the other available agent (competitor in Table 4) was the agent of a depicted action for the sentence verb (Condition 3), and when there was no competitor (Condition 4). When the rated character was not a plausible agent of the verb action (Conditions 1, 2, 5, and 6), ratings were comparatively lower (Table 4). The difference between the plausible Conditions 3 and 4 and each of the other conditions where the rated character was a nonstereotypical agent was significant (repeated

Table 4
Off-line plausibility ratings for the stimuli in Experiment 2

Condition	Stereotypical Agent	Depicted-Action Agent	Competitor	Mean Plausibility Rating
1	no	no	depicted	1.98
2	no	no	stereotypical	1.60
3	yes	no	depicted	5.99
4	yes	no	none	5.86
5	no	yes	stereotypical	1.82
6	no	yes	none	1.92

measures ANOVA, *p*s < .0001; Greenhouse–Geisser adjustments were performed when sphericity was violated; Bonferroni adjustments were used for multiple pairwise comparisons).

Based on the results from the ratings we can be certain that the character that functioned as stereotypical agent in Experiment 2 was a more typical agent of the verb-action than the other, nonstereotypical agent on an image. Even when the action indicated by the verb was depicted (Conditions 5 and 6 in Table 4), plausibility ratings for the agent depicted as performing the action were low, showing that action depiction alone does not render a scene character a more plausible agent.

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 12 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 online comprehension. Such strategy could be induced if participants never encountered a plausible action that was also depicted. Including 12 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 32 participants heard only one of the eight sentences of an item, and the order of appearance of items was randomized individually for every participant. We constructed the lists so that there were no repetitions of item utterances or utterance parts (because verbs and noun phrases were repeated between two items for counterbalancing reasons). The preview time for images was 1,500 msec. An SMI Eye-Link I head-mounted eye tracker monitored participants' eye movements in the depicted scenes with a frequency of 250 Hz while participants were listening to sentences that described part of the scenes. Prior to the experiment, participants were told to listen to the sentences and to inspect the images, and to try to understand both sentences and depicted scenes. There was no other task. The entire experiment lasted approximately 25 min.

3.1.3. Analysis

The time regions for the inferential analyses were the verb (VERB), the postverbal adverb (ADV), and the second noun phrase (NP2). The VERB region stretched from verb onset to the onset of the adverb (mean duration of 1,003 msec for a1, 1,008 msec for a2, 989 msec for b1, and 999 msec for b2; see Table 2). The ADV region extended from the onset of the adverb to the onset of the second noun phrase (mean duration of 1,004 msec for a1, 1,001 msec for a2, 1,013 msec for b1, and 1,002 msec for b2). The NP2 region ranged from the onset of the second noun phrase to its offset (mean duration of 1,056 msec for a1, 1,042 msec for a2, 1,062 msec for b1, and 1,064 msec for b2).

For the inferential analysis of inspection proportions, inspection counts for a time region in the unique (Table 2, a1 and a2) and in the ambiguous identification conditions (Table 2, b1 and b2) were adjusted to factor combinations of target agent (depicted agent, stereotypical agent), target type (depicted, stereotypical), image direction (1, 2), and either participants ($N = 24$) or items ($N = 24$). Recall that image direction was part of the counterbalancing, and refers to whether a target agent was to the left or right of the patient on an image. Image direction was

included in the analyses for completeness and to show that the complex counterbalancing for Experiment 2 was necessary, permitting us to interpret our findings in spite of positional effects on gaze pattern. 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).

3.2. Results and discussion

We report the mean proportions of inspections to the target agents over the course of the utterance (Figs. 14 and 18), and present descriptive and inferential analyses of the proportion of inspections for individual time regions (Figs. 15–17, Figs. 19–21, and Tables 5–8).

Figures 14 and 18 present the time course of the eye movements, displaying mean proportion of inspections to entities in time frames of 250 msec. Fig. 14 displays inspections to the scene entities (depicted agent, stereotypical agent) in the unique identification conditions (Table 2, a1 and a2). Fig. 18 displays inspections to the depicted and to the stereotypical agent in the ambiguous identification conditions (Table 2, b1 and b2). Inspections to the background and the distractors were left out for clarity of presentation, as were inspections to the patient. All of these are, however, displayed in the graphs of the individual time regions (Figs. 15–17 and Figs. 19–21). Tables 5 through 8 show gaze proportions for the two image directions in the unique and ambiguous identification conditions during the adverb region. We only present gaze proportions for the adverb region because this is the only region where image direction

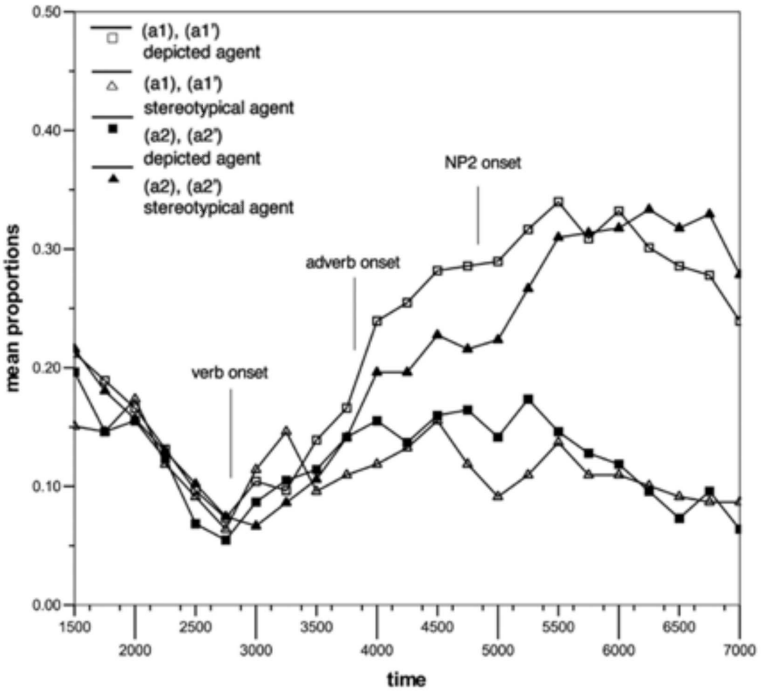


Fig. 14. Time course of the eye movement data for Experiment 2 (unique identification conditions) showing the mean proportion of inspections to target agents from the onset of the spoken stimuli in time frames of 250 msec.

had a significant effect on inspection of the target agent (*p*s > .2 for the VERB and NP2 regions).

3.2.1. Unique identification conditions

The inspection patterns in Fig. 14 show that shortly after adverb onset inspections to the depicted agent increase when the verb identified a depicted target as relevant ('serve-food-to'—detective-serving-food; a1), and at the same time fixations to the stereotypical agent for the depicted target condition are relatively low. In contrast, when the verb identified a stereotypical agent as relevant ('jinx'—wizard; a2), inspections to the stereotypical agent clearly exceeded inspections to the other agent (the detective) during the adverb region. This gaze pattern continues throughout the second noun phrase.

Table 5 shows inspection proportions for the unique identification conditions during the ADV region depending on image direction. There was a preference to inspect the agent to the right more often than the agent to the left of the patient. When the stereotypical agent was left of the patient and the depicted agent was right (Direction 1), people inspected the depicted agent more often than the stereotypical agent. When the stereotypical agent was right of the patient, however, and the depicted agent was leftmost (Direction 2), then participants inspected the stereotypical agent more often than the depicted agent.

Crucially, however, this effect of image direction did not influence the relative inspection proportions to the target agents in the depicted compared with the stereotypical target type condition, which is similar for both directions (see Table 6).

Findings in the unique identification conditions thus clearly replicated the eye gaze patterns from Kamide, Scheepers, and Altmann (2003) and Knoeferle et al. (2005). They demonstrate that—when uniquely identified as relevant—both stereotypical knowledge and depicted infor-

Table 5
Inspection proportions to characters during the ADV region for the unique conditions depending on image direction

Image Direction	Background	Patient	Stereotypical	Depicted	Distractors	Total
1	22.4	13.3	26.5	35.7	2.0	100.0
2	22.2	12.2	38.9	24.4	2.3	100.0

Table 6
Inspection proportions during the ADV region for the unique conditions depending on image direction and target type

Image direction	Target Type	Background	Patient	Stereotypical	Depicted	Distractors	Total
1	Depicted target	21.3	8.8	22.1	44.1	3.7	100.0
	Stereotypical target	23.4	17.1	30.4	28.5	0.6	100.0
2	Depicted target	25.2	12.2	27.2	32.7	2.7	100.0
	Stereotypical target	19.5	12.2	49.4	17.1	1.8	100.0

mation about who does what to whom enable anticipation of the appropriate agent role filler in the scene.

The inferential analyses for the unique identification conditions confirmed the observations for the time course of the eye movements (Figs. 15 to 17 and Tables 5 and 6). During the VERB region, there were no significant main effects for target agent or target type (all *p*s > .1), and no significant interaction of target agent and target type (*p*s > .2; Fig. 15).

Fig. 16 shows the proportion of inspections to scene entities during the ADV region. For the unique identification condition, there was a significant interaction between target agent (depicted agent, stereotypical agent) and target type (depicted target, stereotypical target), $LR\chi^2$(subj) = 17.59, $df = 1$, $p < .0001$; $LR\chi^2$(item) = 26.71, $df = 1$, $p < .0001$. There was no main effect of target type (depicted target, stereotypical target) or target agent (all *p*s > .2).

Log-linear contrast revealed that the interaction of target type and target agent was due to a significantly higher percentage of inspections to the depicted agent for the depicted than stereotypical target condition, $LR\chi^2$(subj) = 16.78, $df = 1$, $p < .0001$; $LR\chi^2$(item) = 21.18, $df = 1$, $p < .0001$. Contrasts for more inspections to the stereotypical agent for the stereotypical compared with the depicted target condition were significant by items, $LR\chi^2$(item) = 18.50, $df = 1$, $p < .001$, but not by participants ($p > .05$).

During the ADV region, we further found a significant interaction of target agent (depicted, stereotypical agent) with image direction (1, 2), $LR\chi^2$(subj) = 12.40, $df = 1$, $p < .001$; $LR\chi^2$(item) = 15.75, $df = 1$, $p < .001$. The interaction resulted from a higher proportion of inspections to the depicted agent for image Direction 1 (right position) than Direction 2 (left position), $LR\chi^2$(subj) = 7.17, $df = 1$, $p < .01$; $LR\chi^2$(item) = 12.02, $df = 1$, $p < .001$. Contrasts for more inspection of the stereotypical character for image Direction 2 (right position) compared

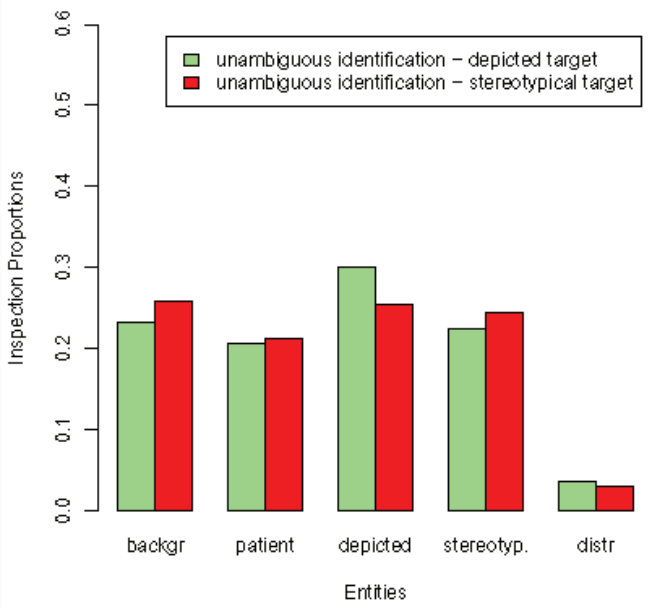


Fig. 15. Percentage of inspections to entities during the VERB region for the unique identification condition in Experiment 2.

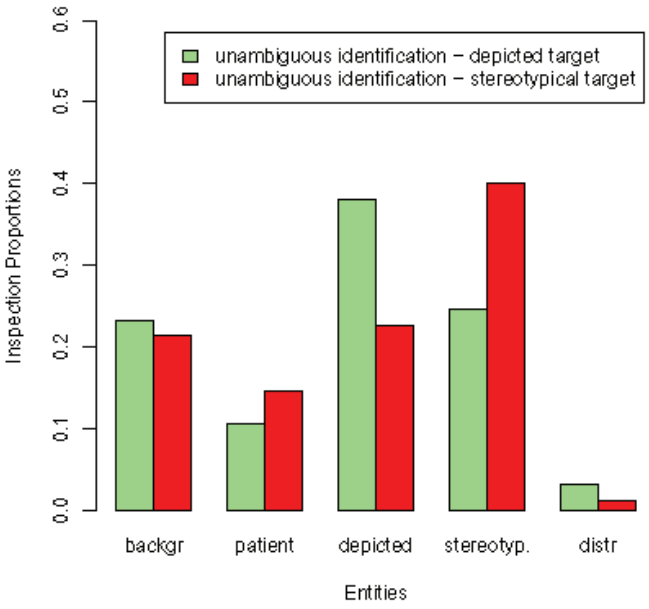


Fig. 16. Percentage of inspections to entities during the ADV region for the unique identification condition in Experiment 2.

with Direction 1 (left position) were significant by items ($p > .09$ by participants), $LR\chi^2$(item) = 13.75, $df = 1$, $p < .001$. There was no significant interaction of target type, target agent, and image direction, $LR\chi^2$s < 1.

For the NP2 region, analyses revealed a significant interaction between target agent (depicted agent, stereotypical agent) and target type (depicted, stereotypical), $LR\chi^2$(subj) = 46.86, $df = 1$, $p < .0001$; $LR\chi^2$(item) = 52.29, $df = 1$, $p < .0001$, in the absence of a main effect of target agent, $LR\chi^2$s < 1 (Fig. 17). Contrasts showed that the interaction was due to a higher proportion of inspections to the depicted agent for the depicted than stereotypical target type, $LR\chi^2$(subj) = 4.82, $df = 1$, $p < .05$; $p > .1$ by items, and to a higher proportion of inspections to the stereotypical than to the depicted agent for the stereotypical target type sentences, $LR\chi^2$(subj) = 851.72, $df = 1$, $p < .0001$; $LR\chi^2$(item) = 851.72, $df = 1$, $p < .0001$.

3.2.2. Ambiguous identification conditions

Following the onset of the adverb, we observed more anticipatory fixations to the agent of the depicted spying event (the wizard), than to the stereotypical agent (the detective) for sentences b1 and b2 as shown in Fig. 18. Note that no early interaction was expected for conditions b1 and b2 during the adverb time region (Table 2) because the b1 and b2 stimuli were identical prior to the second noun phrase. Rather, we expected a main effect with the interesting issue being the directionality of the effect. The anticipatory looks to the depicted agent for b1 and b2 occurred after people had heard *Den Piloten (ACC) bepitzelt …* ('The pilot (ACC) spies on … ') and crucially before they heard the respective second noun, which then disambiguated toward the depicted (b1) or stereotypical (b2) target type. The comprehension system thus had to

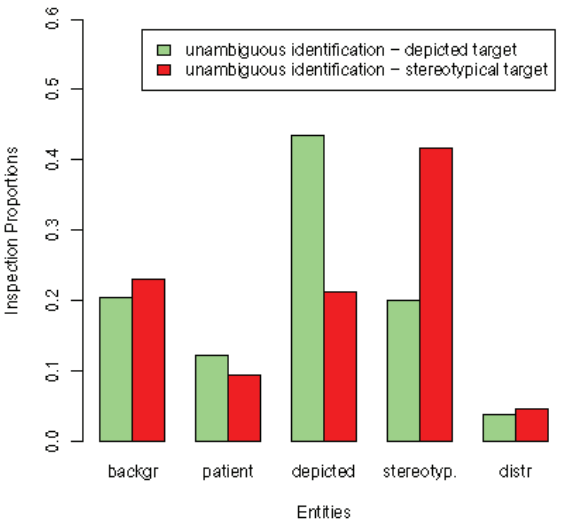


Fig. 17. Percentage of inspections to entities during the NP2 region for the unique identification condition in Experiment 2.

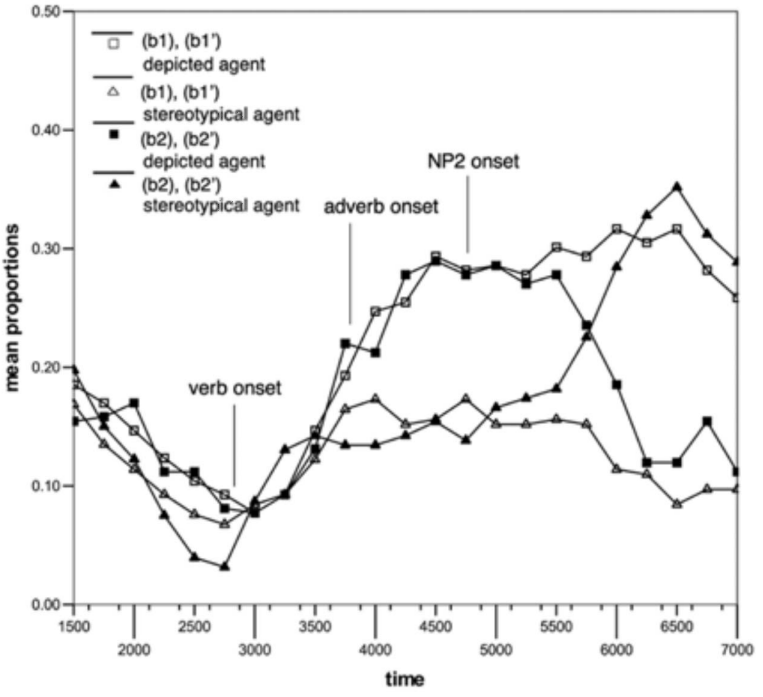


Fig. 18. Time course of the eye movement data for Experiment 2 (ambiguous identification conditions) showing the mean proportion of inspections to the target agents from the onset of the spoken stimuli in time frames of 250 msec.

revise the interpretation of the utterance for the stereotypical target condition (b2) during the second noun phrase. Indeed, the eye movement patterns provide evidence for revision of the initial depicted-event preference. Late during the NP2 region, inspections to the stereotypical agent for the stereotypical target condition (b2) increase, while looks to the depicted agent for this condition decrease. After this disambiguation the pattern of inspections is the same as for the unique identification conditions (see Figs. 14 and 18).

Table 7 displays inspection proportions to the characters for the adverb region depending on image direction. There was a tendency toward more inspections to the depicted agent when it was to the right (Direction 2) than to the left of the patient (Direction 1), and more gazes to the stereotypical agent when it was to the right (Direction 1) than to the left of the patient (Direction 2). Overall, however, the gaze pattern to the depicted and stereotypical agents was similar for both image directions, revealing more inspection of the depicted than stereotypical agent for both image directions (Table 7) and in the depicted as well as the stereotypical target conditions (Table 8).

Inferential analyses for the ambiguous identification conditions confirmed these observations. The main effect of direction; the interaction between image direction and target agent; interaction between image direction and target type; and interaction among image direction, target agent, and target type were nonsignificant for all the analysis regions ($ps > .2$).

For the VERB region, there was no significant main effect of target agent (depicted agent vs. stereotypical agent) or target type (depicted, stereotypical), $LR\chi^2s < 1$, nor was there a significant interaction of these two factors, $LR\chi^2s < 1$ (Fig. 19).

Fig. 20 shows the proportions of inspections to entities during the ADV region for the ambiguous identification condition. We found a significant main effect of target agent (depicted

Table 7
Inspection proportions to characters during the ADV region for the ambiguous conditions depending on image direction

Image Direction	Background	Patient	Stereotypical	Depicted	Distractors	Total
1	17.3	10.9	30.6	39.1	2.0	100.0
2	18.2	7.1	21.5	51.9	1.3	100.0

Table 8
Inspection proportions to characters during the ADV region for the ambiguous conditions depending on image direction and target type

Image Direction	Target Type	Background	Patient	Stereotypical	Depicted	Distractors	Total
1	Depicted target	20.5	11.3	27.8	37.7	2.6	100.0
	Stereotypical target	14.0	10.5	33.6	40.6	1.4	100.0
2	Depicted target	16.1	6.6	21.9	54.0	1.5	100.0
	Stereotypical target	20.0	7.5	21.3	50.0	1.3	100.0

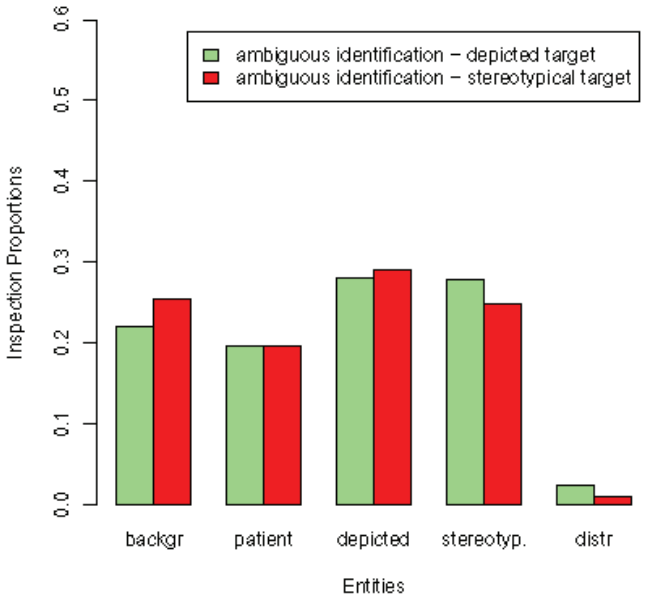


Fig. 19. Percentage of inspections to entities during the VERB region for the ambiguous identification condition in Experiment 2.

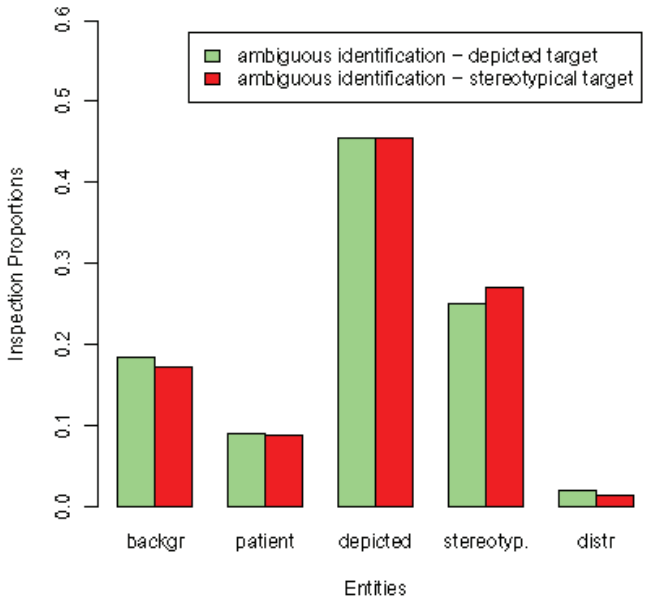


Fig. 20. Percentage of inspections to entities during the ADV region for the ambiguous identification condition in Experiment 2.

agent, stereotypical agent), $LR\chi^2$(subj) = 31.66, df = 1, p < .0001; $LR\chi^2$(item) = 31.66, df = 1, p < .0001, in the absence of an interaction of target agent and identification, $LR\chi^2$s < 1.

The difference between the main effect for the ambiguous identification condition (b1 and b2), and the reliable interaction for the unique identification condition (a1 and a2, Table 2) was significant for the ADV region. Analyses revealed a three-way interaction among target agent (depicted agent, stereotypical agent), identification (unique, ambiguous) and target type (depicted target, stereotypical target), $LR\chi^2$(subj) = 8.77, df = 1, p < .01; $LR\chi^2$(item) = 11.21, df = 1, p < .001.

For the NP2 region in the ambiguous identification condition (Fig. 21), log-linear analyses confirmed that the interaction between target agent and identification was not significant (p > .1). However, when extending the NP2 region until the end of the utterance, the interaction turned significant. For this later NP2 region (from NP2 onset to the end of the utterance), there was an interaction between target agent (depicted agent, stereotypical agent) and identification (depicted, stereotypical), $LR\chi^2$(subj) = 61.65, df = 1, p < .0001; $LR\chi^2$(item) = 61.36, df = 1, p < .0001. Contrasts revealed a significantly higher proportion of inspections to the depicted than to the stereotypical agent for the depicted target condition, $LR\chi^2$(subj) = 30.26, df = 1, p < .0001; $LR\chi^2$(item) = 27.14, df = 1, p < .0001, and a significantly higher percentage of inspections to the stereotypical in comparison with the depicted agent for the stereotypical target type condition, $LR\chi^2$(subj) = 68.28, df = 1, p < .001; $LR\chi^2$(item) = 69.17, df = 1, p < .001.

The central finding of Experiment 2 is the preference of comprehension processes to rely on depicted actions and their associated role relations for early thematic role assignment when the verb identified both thematic relations established by stored knowledge and role relations pro-

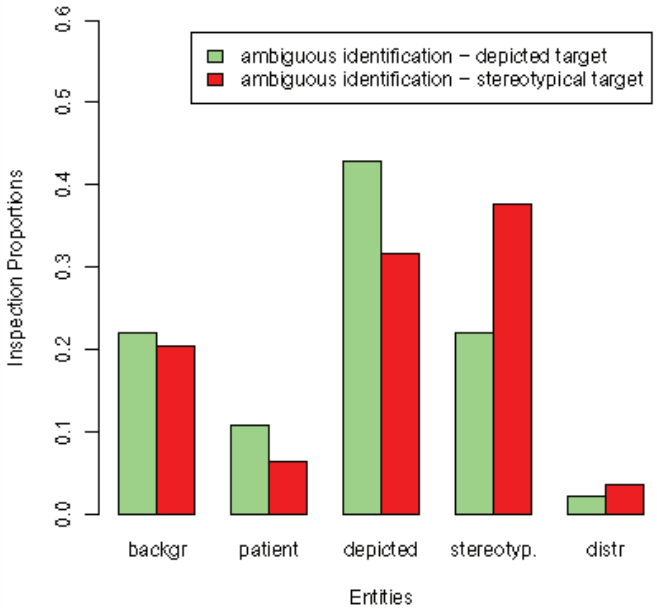


Fig. 21. Percentage of inspections to entities during the NP2 region for the ambiguous identification condition in Experiment 2.

vided by the scene events as relevant (Fig. 18). Clearly, this finding is only meaningful in comparison with the early disambiguation toward either a depicted or a stereotypical agent in the unique identification condition when the verb identified either depicted role relations or stereotypical thematic relations as relevant for comprehension (Fig. 14).

Although the position of an agent affected gaze pattern during the adverb region, the relative proportion of inspections to the target agents depending on target type are similar for both directions. This together with the significant interaction between target agent and target type for the unique, and the main effect of depicted agent in the ambiguous identification conditions suggest that our findings hold true independent of image direction, which was counterbalanced in our design.

3.2.3. Analyses by blocks

In addition to the preceding analyses, we carried out an analysis by blocks (first vs. second half of experiment). This was done to ensure that the observed relative priority of depicted events compared with stereotypical thematic role knowledge did not result from learning strategies over the course of the experiment but rather reflected a general comprehension mechanism.

Tables 9 and 10 give an overview of the inspection proportions to characters for the main analysis regions (adverb and second noun phrase) for the first and second block of the experiment. We start by detailing the results for the adverb region (Table 9) because it was gaze pattern during this region that supported the relative priority of depicted events. Crucially, Table 9 shows the same gaze pattern for conditions in the first compared with the second block for the adverb region. Gaze patterns in each block confirm our findings of an interaction between target agent and target type for the unique, and a main effect of target agent for the ambiguous identification conditions.

Log-linear analyses in the unique ($p > .1$ by participants and $LR\chi^2 < 1$ by items) and in the ambiguous identification conditions ($LR\chi^2$s < 1) showed that there was no significant interaction among target agent (depicted agent, stereotypical agent), target type, and block (first, second).

For the second noun phrase, Table 10 shows the same overall gaze pattern for all conditions when comparing the first and second block. Log-linear analyses for the unique identification conditions revealed a significant interaction of target agent (depicted agent, stereotypical agent), target type, and block (first, second), $LR\chi^2$(subj) = 6.16, $df = 1$, $p < .05$; $LR\chi^2$(item) = 7.87, $df = 1$, $p < .01$. No interaction among target agent (depicted agent, stereotypical agent), target type, and block (first, second) was found for the ambiguous identification conditions ($LR\chi^2$s < 1).

The interaction for the unique identification conditions together with the counterbalancing, the plausibility ratings, and the analyses by blocks support the straightforward interpretability of the preference for depicted events over verb-based thematic knowledge in thematic interpretation. At the same time, findings from Experiment 2 further confirmed the tight temporal coordination between when a verb identifies a relevant depicted event, and when that event influences comprehension. The implications of these findings for an account of visually situated language comprehension are considered next.

Table 9
Inspection proportions to characters during the ADV region for Experiment 2 depending on block

Condition	Block		Background	Patient	Stereotypical	Depicted	Distractors	Total
Unique identification	1	Depicted target	21.2	9.8	27.3	40.2	1.5	100.0
		Stereotypical target	22.8	18.1	39.2	18.1	1.8	100.0
Unique identification	2	Depicted target	25.2	11.3	22.5	36.4	4.6	100.0
		Stereotypical target	19.9	10.6	41.1	27.8	0.7	100.0
Ambiguous identification	1	Depicted target	18.7	9.4	18.7	51.1	2.2	100.0
		Stereotypical. target	16.1	5.6	22.4	53.8	2.1	100.0
Ambiguous identification	2	Depicted target	18.1	8.7	30.9	40.3	2.0	100.0
		Stereotypical target	18.1	11.9	31.3	38.1	0.6	100.0

Table 10
Inspection proportions to characters during the NP2 region for Experiment 2 depending on block

Condition	Block		Background	Patient	Stereotypical	Depicted	Distractors	Total
Unique identification	1	Depicted target	21.5	15.6	22.5	36.4	4.0	100.0
		Stereotypical target	22.4	12.7	40.2	19.0	5.7	100.0
Unique identification	2	Depicted target	20.3	15.2	14.8	43.9	5.8	100.0
		Stereotypical target	22.6	10.4	42.4	17.4	7.3	100.0
Ambiguous identification	1	Depicted target	18.1	13.7	21.1	42.5	4.7	100.0
		Stereotypical target	19.6	7.1	41.2	25.1	7.1	100.0
Ambiguous identification	2	Depicted target	23.1	11.4	22.2	39.6	3.6	100.0
		Stereotypical target	18.2	7.0	45.4	28.1	1.3	100.0

4. General discussion

These experiments investigated the temporal and informational dimensions of situated utterance comprehension. Prior studies of situated utterance comprehension and insights from developmental research have given us reason to expect a tight temporal coordination among adult utterance comprehension, attention in the scene, and the verb-mediated influence of depicted events on comprehension, as well as a great—albeit not necessarily primary—importance of verb-mediated depicted events over stereotypical thematic knowledge.

Our findings confirmed these expectations, revealing two important characteristics of what we dub the coordinated interplay of situated utterance comprehension: First, they demonstrate a close temporal coordination between when relevant depicted events are identified by the utterance, and the point in time—relative to their identification as likely targets—at which those depicted events enable thematic role assignment and structural disambiguation. Second, our data indicate a greater relative importance of verb-mediated depicted events than verb-based thematic role knowledge for incremental thematic interpretation in situated utterance comprehension.

Experiment 1a provides strong evidence for the rapid, verb-mediated effect of depicted agent–action–patient events on incremental thematic role assignment and structural disambiguation of initially ambiguous MV/RR clause sentences. Shortly after the verb, anticipatory eye movements to the patient and agent reveal the effects of depicted events on incremental thematic role assignment and structural disambiguation of initially ambiguous main clause and RR clause sentences, respectively. The claim that verb-mediated depicted events do indeed trigger disambiguation of local structural ambiguity is corroborated by data from eye tracking studies that directly compared comprehension of initially structurally ambiguous with unambiguous sentences (Knoeferle, in press).

Findings from Experiment 1b suggest that indirect tense references to a relevant event can also rapidly make depicted events available for incremental thematic role assignment. Disambiguation was evident in the gaze pattern during the verb in the ambiguous clause, and thus earlier than when no tense cues preceded the verb (see Experiment 1a, and Experiments 1 and 2 in Knoeferle et al., 2005). These gaze patterns could, however, either be triggered by the tense cues alone, or these cues may simply facilitate use of the verb. Following developmental research that suggests a greater importance of the immediate scene for the acquisition of concrete than abstract words (e.g., Gillette et al., 1999), we expect the temporal coordination of utterance comprehension with scene processing in adult comprehension will be most robust for words that directly refer to objects and events in the scene.

Experiment 2 demonstrated that stored knowledge of thematic role relations (that were not depicted) and immediately depicted events (that were nonstereotypical) both contribute to rapid thematic role interpretation. When the utterance uniquely identified an agent as relevant by means of the action he performed (no other agent in the scene was a plausible agent given the verb, or depicted as performing the verb action), we found evidence for early incremental thematic interpretation shortly after the verb, based on the depicted action. When the verb identified an agent as relevant on the basis of stored thematic role knowledge (no other agent in the scene was a plausible agent or depicted as performing the action indicated by the verb), gaze pattern also revealed early incremental thematic interpretation using verb-based stereotypical knowledge.

In contrast, when the utterance did not determine uniquely whether the comprehension system should rely on stored thematic role knowledge (identifying a relevant stereotypical agent) or on depicted events (identifying an alternative agent of a depicted event as relevant), we observed a strong preference of the comprehension system to rapidly rely on depicted events over stored thematic knowledge for processes of incremental thematic role assignment. Evidence for this came from a higher proportion of anticipatory eye movements to the agent depicted as performing the verb action in comparison with the agent that was stereotypical for the verb.

The findings from Experiment 1 extend previous research on the coordination of utterance and scene processing. Tanenhaus et al. (1995) demonstrated that the type of visual referential context biases people to either adopt a destination (one-referent context) or a location interpretation (two-referent context) of a temporarily structurally ambiguous phrase. We argued that the gaze pattern in their studies was compatible with two alternative interpretations. Either the referential visual context influenced structuring of the utterance asynchronously of comprehension, or its use was triggered by the utterance (i.e., resulting from people hearing *the apple*). The fact that the pattern of eye gaze between the two types of contexts differs from the very start of the utterance makes it problematic to tease apart these two interpretations, especially as the scene itself varies in the two conditions.

For our investigation, in contrast, determining when the scene events influenced comprehension relative to when the utterance identified them as relevant was possible because the scene in Experiment 1 was kept constant (unlike in Tanenhaus et al., 1995), and rather the meaning of the verb in the utterance varied. Eye movement patterns in Experiment 1a did not differ between the MV and RR clause condition prior to the offset of the verb. It was only after people had heard the verb that one of the two available actions was identified as relevant and that role relations established by that action rapidly constrained utterance interpretation. Gaze pattern in Experiment 1b revealed a less robust temporal coordination for indirect tense cues. Findings from Experiments 1a and 1b are thus compatible with an account that prioritizes the important function of direct referential expressions such as lexical verbs in making depicted actions rapidly available for comprehension.

Findings from Experiments 1 and 2 extend results by Kamide, Scheepers, and Altmann (2003) and Knoeferle et al. (2005) that showed the rapid influence of either linguistic/world knowledge or depicted events on thematic role assignment, respectively. Experiment 1 generalizes findings on German SVO/OVS from Knoeferle et al. (2005) to another language (English) and sentence structure (MV/RR clause). Experiment 2 adds the insight that listeners prefer to rely on verb-mediated depicted events over stereotypical knowledge of events for thematic role assignment when the verb identifies both of these informational sources as relevant for utterance comprehension.

The findings from Experiments 1 and 2 together with prior research on situated utterance comprehension, and insights from the acquisition literature, prompt us to propose the *coordinated interplay account* (CIA) of situated utterance comprehension. The CIA identifies two fundamental steps in situated utterance comprehension. First, comprehension of the unfolding utterance guides attention in the scene, establishing reference to objects and events (Tanenhaus et al., 1995; Knoeferle et al., 2005), and anticipating likely referents (see Altmann & Kamide, 1999). Once the utterance has identified the most likely object or event, and attention has shifted to it, the attended scene information then rapidly influences utterance comprehension

(see Experiment 1a). The CIA further assumes that the close time-lock between utterance comprehension and attention in the scene (e.g., Tanenhaus et al., 1995) involves a strategy of first checking the scene rather than solely relying on linguistic/world knowledge. Such a strategy might lead to the greater relative priority of immediately depicted events over knowledge of stereotypical events in comprehension that we observed in Experiment 2.

Findings from both our experiments and the CIA are broadly compatible with various interactionist and embodied accounts of online language processing (e.g., Barsalou, 1999b; Bergen & Chang, 2005; Chambers et al., 2004; Jackendoff, 2002; Tanenhaus, Spivey-Knowlton, & Hanna, 2000; Zwaan, 2004). The important contribution of our results is the first step toward a processing account of situated utterance comprehension. Although findings from our experiments are compatible with these situated frameworks, they are not obviously predicted by them, as these frameworks are not yet sufficiently specified to make concrete predictions about the temporal and informational dimensions of situated utterance comprehension. Psycholinguistic theories of comprehension, in contrast (e.g., Frazier & Clifton, 1996; Pickering et al., 2000; Townsend & Bever, 2001; Trueswell & Tanenhaus, 1994) provide a detailed account of when which informational source can be used. Because they do not, however, explicitly include scene information, it is difficult to derive concrete predictions on situated utterance comprehension from them. Models such as the competitive integration model (e.g., Spivey-Knowlton, 1994; Tanenhaus et al., 2000) or the constraint-based interactionist model by MacDonald et al. (1994) similarly make no clear prediction for the research questions examined by Experiments 1 and 2. These accounts were designed to model ambiguity resolution when there was only linguistic/world knowledge available, not, however, visual scenes. It is further not clear what they would predict for incremental thematic interpretation in visual scenes (i.e., when there is no structural ambiguity) as in Experiment 2.

Recent modeling using a simple recurrent connectionist network (see Elman, 1990) has, however, shown that an interactionist approach is generally well suited to modeling situated utterance comprehension (Mayberry, Crocker, & Knoeferle, 2005). Their network succeeds in modeling the findings by Kamide, Scheepers, and Altmann (2003) and Knoeferle et al. (2005), showing the rapid incremental use of either linguistic and verb-based world knowledge or of depicted events. In a follow-on simulation, it moreover developed the behavior that we observed in people for Experiment 2—a greater relative priority of the immediate events over stereotypical thematic role knowledge in thematic role assignment shortly after the verb is encountered. A further computational model of spoken language comprehension suitable for modeling our findings appears to be Fuse by Roy and Mukherjee (2005). It includes a dynamic model of visual attention that enables anticipating the most likely objects in a scene based on processing of the unfolding utterance. Among the embodied frameworks, the model by Bergen and Chang (2005), for example, would also seem to be well suited for modeling a closely synchronized interaction between utterance and scene processing due to its use of executive schemas that allow modeling sequential, concurrent, and asynchronous events.

In addition to being compatible with the preceding comprehension frameworks and models, our findings are also compatible—even expected—based on insights from language acquisition. It has been shown that the immediate environment had a greater influence on the acquisition of words when child-directed speech about an object coincided with the child's attention to that object than when it did not (e.g., Dunham et al., 1993; Harris et al., 1986). Assuming

that there is at least a partial acquisition–comprehension continuum, the obvious prediction for adult language comprehension is that the first step of the CIA (utterance-mediated attention to an object or event) triggers the rapid use of that scene information in comprehension. The tight coordination of utterance comprehension and attention in a scene together with the importance of scene information for the acquisition of concrete words (e.g., Gillette et al., 1999), further suggest that the comprehension system is geared toward first checking for information in the scene rather than solely relying on linguistic and world knowledge for utterance comprehension.

Findings from studies that tracked children's eye movements in scenes during comprehension (e.g., Snedeker & Trueswell, 2004; Trueswell et al., 1999) at first sight appear to clash with the ideas from acquisition research. Trueswell et al. (1999) reported that unlike adults, 5-year-olds are not able to rapidly use a visual referential context for the online structuring of an utterance. Snedeker and Trueswell (2004) further extended this research, finding that whereas adults rely on lexical and referential information for syntactic structuring of an utterance, children only rely on lexical (verb bias) information. Snedeker and Trueswell further reported that eye movements in the scene revealed the beginnings of sensitivity to referential context.

We think that their data can be attributed to the type of information in the scene relevant to structuring of the utterance, and the way in which that scene information was identified as relevant by the utterance. The CIA predicts a tight temporal coordination between utterance comprehension, attention to likely objects or events, and the influence of information about those objects or events on comprehension for direct reference between a word and a scene object or event (e.g., verb action). In studies by Trueswell et al. (1999) and Snedeker and Trueswell (2004), no such direct cue in the utterance uniquely made the relevant scene information available. Rather, to use the referential context established by the scenes in these studies, children must establish a contrastive relation between two scene objects. In contrast, for depicted events, relations between scene objects are explicitly depicted by the actions (see Knoeferle et al., 2005), and are thus directly available rather than first requiring contrastive comparison. Owing to these differences, we would expect that children at the age of 5 will be able to use verb-mediated depicted agent–action–patient events for rapid thematic role assignment and structural disambiguation.

The preceding discussion emphasizes the fundamental role that the CIA accords to the utterance in identifying those objects or events that are most informative for comprehension. Although our account emphasizes a tight temporal coordination of utterance comprehension and attention, perceptual processes can extract information from a scene prior to utterance mediation. Support for the hypothesis that we interpret the visual environment in the process of perception comes from research on Gestalt psychology (e.g., Koehler, 1947; Wertheimer, 1923/1938; see Knoeferle et al., 2005, for discussion). Our findings are compatible with this position. They add the insight that interpretation of visual information in perception leads to the recovery of propositions rapidly enough to influence disambiguation and thematic role-assignment processes online once a concrete word in the utterance identifies which objects and events—in the possibly complex scene—are relevant for comprehension. In increasingly complex settings, such a guiding role of the utterance may become even more important for guiding attention to informative objects and events. Moreover, once the utterance has iden-

tified scene information as relevant, other objects and events that are related to that scene information can also inform comprehension.

The various experimental situations investigating the interplay between visual processing and comprehension that we have discussed share the characteristic that the utterances are about the immediate visual scene. In language acquisition, when parents talk to their children, language is likely often about entities and events in the immediate environment that children typically explore (e.g., Snow, 1977). In these situations, it serves a specific function, namely making the immediate scene accessible for the child, and identifying objects that are relevant for comprehension and acquisition (see, e.g., Richards & Goldfarb, 1986; Roy & Pentland, 2002; Siskind, 1990, for related modeling research).

During adult language comprehension, however, this is not true to the same extent. Language often subserves other functions or tasks and is only used to refer to entities in the immediate scene for part of the communication in which we engage. Much of our day is spent talking, reading, or writing about things that are not immediately present.

The CIA is an account of situated utterance comprehension. In situations where the utterance does not directly relate to the immediate visual environment, immediately depicted events will almost certainly not have the importance that is suggested by findings from Experiments 1 and 2 in this article because the scene is irrelevant. It is the immediate presence and relevance of utterance, linguistic/world knowledge, and depicted events for comprehension that enables the rapid interplay between these informational sources. We do expect, however, that in situations where the utterance is about the immediate environment, our findings of the rapid, verb-mediated influence of depicted events on structural disambiguation, and of the priority of depicted events over verb-based thematic role knowledge in thematic interpretation will apply.

Although language is often not about the immediate scene in adult life, we have spent a substantial part of our lives acquiring language. We suggest that this period may indeed have shaped both our cognitive architecture, resulting in a rapid interaction between cognitive systems such as language and vision, and comprehension mechanisms, enabling us to rapidly avail ourselves of explicit role relations from the immediate scene when the utterance identifies them as relevant. This might, in particular, hold true when the scene information in question is "fleeting," as is the case with action events in real life. This time-limited presence of action events in comparison with the virtually continuous presence of our knowledge might contribute to the rapid use of verb-mediated depicted events, and to the greater relative priority of depicted events over stereotypical thematic role knowledge during real-time comprehension.

Acknowledgments

We thank Christoph Scheepers for helpful comments on the first experiment, and for access to the eye tracking laboratory of the Department of Psychology at the University of Dundee (U.K.). We are grateful to Martin Pickering for comments on an early draft of this article. This research was funded by a PhD scholarship to Pia Knoeferle and by SFB 378"ALPHA" to Matthew W. Crocker, both awarded by the German research foundation (DFG).

References

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.

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

Altmann, G. T. M., & Steedman, M. (1988). Interaction with context during human sentence processing. *Cognition, 30,* 191–238.

Baayen, R. H., Pipenbrock, R., & Gulikers, L. (1995). *The celex lexical database* [CD-ROM]. Philadelphia: University of Pennsylvania, Linguistic Data Consortium.

Barsalou, L. W. (1999a). Language comprehension: Archival memory or preparation for situated action? *Discourse Processes, 28,* 61–80.

Barsalou, L. W. (1999b). Perceptual and symbol systems. *Behavioral and Brain Sciences, 22,* 577–609.

Bergen, B., & Chang, N. (2005). Embodied construction grammar in simulation-based language understanding. In J.-O. Östman & M. Fried (Eds.), *Construction grammar(s): Cognitive and cross-language dimensions* (pp. 147–190). Amsterdam: John Benjamins.

Bergen, B., Chang, N., & Narayan, S. (2004). Simulated action in an embodied construction grammar. In K. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the 26th Annual Conference of the Cognitive Science Society* (pp. 108–113). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Bever, T. (1970). The cognitive basis for linguistic structures. In J. R. Hayes (Ed.), *Cognition and the development of language* (pp. 279–362). New York: Wiley.

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

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.

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

Crain, S., & Steedman, M. (1985). On not being led up the garden path: The use of context by the psychological parser. In D. Dowty, L. Karttunnen, & A. Zwicky (Eds.), *Natural language parsing* (pp. 320–358). Cambridge, MA: Cambridge University Press.

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

Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14,* 179–211.

Frazier, L., & Clifton, C. (1996). *Construal.* Cambridge, MA: MIT Press.

Gillette, J., Gleitman, H., Gleitman, L., & Lederer, A. (1999). Human simulations of vocabulary learning. *Cognition, 73,* 135–176.

Gleitman, L. (1990). The structural sources of verb meanings. *Language Acquisition, 1,* 3–55.

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.

Hemforth, B. (1993). *Kognitives Parsing: Repräsentation und Verarbeitung sprachlichen Wissens* [Cognitive parsing: Representation and processing of linguistic knowledge]. Sankt Augustin, Germany: Infix-Verlag.

Howell, D. C. (2002). *Statistical methods for psychology* (5th ed.). Pacific Grove, CA: 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., Altmann, G. T. M., & Haywood, S. (2003). The time course of prediction in incremental sentence processing: Evidence from anticipatory eye-movements. *Journal of Memory and Language, 49,* 133–156.

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 movements: A window on mind & brain.* Oxford, England: Elsevier.

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.

Koehler, W. (1947). *Gestalt psychology.* New York: Liveright.

Macdonald, M. C. (1994). Probabilistic constraints and syntactic ambiguity resolution. *Language and Cognitive Processes, 9,* 157–201.

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review, 101,* 676–703.

Matin, E., Shao, K., & Boff, K. (1993). Saccadic overhead: Information processing time with and without saccades. *Perceptual Psychophysics, 53,* 372–380.

Mayberry, M., Crocker, M. W., & Knoeferle, P. (2005). A connectionist model of sentence comprehension in visual words. In B. G. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the 27th Annual Conference of the Cognitive Science Society* (pp. 1437–1442). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

McRae, K., Spivey-Knowlton, M. J., & Tanenhaus, M. K. (1998). Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language, 38,* 283–312.

Pickering, M. J., Traxler, M. J., & Crocker, M. W. (2000). Ambiguity resolution in sentence processing: Evidence against frequency-based accounts. *Journal of Memory and Language, 43,* 447–475.

Richards, D., & Goldfarb, J. (1986). The episodic memory model of conceptual development: An integrative viewpoint. *Cognitive Development, 1,* 183–219.

Roy, D., & Mukherjee, N. (2005). Towards situated speech understanding: Visual context priming of language models. *Computer Speech and Language, 19,* 227–248.

Roy, D., & Pentland, A. (2002). Learning words from sights and sounds: A computational model. *Cognitive Science, 26,* 113–146.

Runner, T. R., Sussman, R. S., & Tanenhaus, M. K. (2003). Assignment of reference to reflexives and pronouns in picture noun phrases: Evidence from eye-movements. *Cognition, 89,* B1–B13.

Sedivy, J. C. (2002). Invoking discourse-based contrast sets and resolving syntactic ambiguities. *Journal of Memory and Language, 46,* 341–370.

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

Siskind, J. M. (1990). Acquiring core meanings of words, represented as Jackendoff-style conceptual structures, from correlated streams of linguistic and non-linguistic input. In R. C. Berwick (Ed.), *Proceedings of the 28th Conference of the Association for Computational Linguistics* (pp. 143–156). Pittsburgh, PA: Association for Computational Linguistics.

Snedeker, J., & Trueswell, J. C. (2004). The developing constraints on parsing decisions: The role of lexical-biases and referential scenes in child and adult sentence processing. *Cognitive Psychology, 49,* 238–299.

Snow, C. (1977). Mothers' speech research: From input to interaction. In C. Snow & C. A. Ferguson (Eds.), *Talking to children: Language input and acquisition* (pp. 31–49). 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., Tanenhaus, M. K., Eberhard, K. M., & Sedivy, J. C. (2002). Eye-movements and spoken language comprehension: Effects of visual context on syntactic ambiguity resolution. *Cognitive Psychology, 45,* 447–481.

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

Spivey-Knowlton, M. J. (1994). Quantitative predictions from a constraint-based theory of syntactic ambiguity resolution. In M. Mozer, J. Elman, P. Smolensky, D. Touretzky, & A. Weigand (Eds.), *The 1993 Connectionist Models Summer School* (pp. 130–137). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Sussman, S. R., & Sedivy, J. C. (2003). The time-course of processing syntactic dependencies: Evidence from eye-movements. *Language and Cognitive Processes, 18,* 143–163.

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.

Tanenhaus, M. K., Spivey-Knowlton, M. J., & Hanna, J. E. (2000). Modeling thematic and discourse context effects with a multiple constraints approach: Implications for the architecture of the language comprehension system. In M. W. Crocker, M. J. Pickering, & C. Clifton (Eds.), *Architectures and mechanism for language processing* (pp. 90–118). Cambridge, UK: Cambridge University Press.

Townsend, D. J., & Bever, T. G. (2001). *Sentence comprehension: The integration of habits and rules.* Cambridge, MA: MIT Press.

Trueswell, J. C. (1996). The role of lexical frequency in syntactic ambiguity resolution. *Journal of Memory and Language, 35,* 566–585.

Trueswell, J. C., Sekerina, I., Hill, N., & Logrip, M. (1999). The kindergarten-path effect: Studying online sentence processing in young children. *Cognition, 73,* 89–134.

Trueswell, J. C., & Tanenhaus, M. K. (1994). Toward a lexicalist framework for constraint-based syntactic ambiguity resolution. In C. Clifton, L. Frazier, & K. Rayner (Eds.), *Perspectives in sentence processing* (pp. 155–179). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Trueswell, J. C., Tanenhaus, M. K., & Garnsey, S. M. (1994). Semantic influences on parsing: Use of thematic role information in syntactic disambiguation. *Journal of Memory and Language, 33,* 285–318.

Wertheimer, M. (1938). Laws of organization in perceptual forms. In W. Ellis (Ed.), *A source book of Gestalt Psychology* (pp. 71–80). London: Routledge & Kegan Paul. (Original work published 1923)

Zwaan, R. A. (2004). The immersed experiencer: Towards an embodied theory of language comprehension. In B. H. Ross (Ed.), *The psychology of learning and motivation* (Vol. 44, pp. 35–62). New York: Academic.