The influence of the immediate visual context on incremental thematic role-assignment: evidence from eye-movements in depicted events

Pia Knoeferle

Department of Computational Linguistics, Saarland University, Germany

Matthew W. Crocker

Department of Computational Linguistics, Saarland University, Germany

Christoph Scheepers

Department of Psychology, University of Dundee, Scotland, UK

Martin J. Pickering

Department of Psychology, University of Edinburgh, Scotland, UK

Correspondence concerning this article should be addressed to:

Pia Knoeferle

Department of Computational Linguistics

Saarland University

66041 Saarbrücken

Germany

email: knoeferle@coli.uni-sb.de

telephone: +49(0)681 302 6554

fax: +49 (0)681 302 6551

Abstract

Studies monitoring eye-movements in scenes containing entities have provided robust evidence for incremental reference resolution processes. This paper addresses the less studied question of whether depicted event scenes can affect processes of incremental thematic role-assignment. In Experiments 1 and 2 participants inspected agent-action-patient events while listening to German verb-second sentences with initial structural and role ambiguity. The experiments investigated the time course with which listeners could resolve this ambiguity by relating the verb to the depicted events. Such verb-mediated *visual* event information allowed early disambiguation on-line, as evidenced by anticipatory eye-movements to the appropriate agent/patient role filler. We replicated this finding while investigating the effects of intonation. Experiment 3 demonstrated that when the verb was sentence-final and thus did not establish early reference to the depicted events, linguistic cues alone enabled disambiguation before people encountered the verb. Our results reveal the on-line influence of depicted events on incremental thematic role-assignment and disambiguation of local structural and role ambiguity. In consequence, our findings require a notion of *reference* that includes actions and events in addition to entities (e.g., Jackendoff, 1983), and argue for a theory of on-line sentence comprehension that exploits a rich inventory of semantic categories.

Keywords: language comprehension, eye-movements, depicted events, incremental thematic role-assignment, structural disambiguation, ontological categories, reference

Introduction

It has been stated that an essential property of language is that it refers to things in the world. It has further been demonstrated that in on-line comprehension reference to entities is established rapidly, and that referentially relevant non-linguistic information influences how the linguistic input is structured (Tanenhaus, Spivey-Knowlton, Eberhard & Sedivy, 1995). These insights and results emphasize important aspects of the relation between linguistic representations, the language processing system, and the visual environment. The experiments presented in this article extend these findings in two important respects.

First, we suggest that just as language refers to entities in the immediate visual context, it also refers to other ontological categories such as actions or events. While the definition of the term *reference* has generally only been extended to cover reference from nouns to the ontological category of things (Crystal, 2003), the experiments reported in this paper are compatible with a notion of reference that includes reference from linguistic expressions like verbs and sentences to a broad variety of corresponding distinct ontological categories such as actions and events (e.g., Jackendoff, 1983). We propose reference to these ontological categories can be established on-line.

Second, we argue that visual scenes crucially can reveal important relations between things, in addition to identifying entities themselves. Through these relations, the visual context may reveal an entity's role with respect to other objects. It is suggested events are one ontological category that provides visual information about role-relations between entities. Just as referentially relevant non-linguistic information has been found to influence the time course of referential processes, we propose that non-linguistic information that establishes role-relations between entities will affect incremental thematic role assignment. The present paper directly investigates the extent to which depicted events, which identify such relevant relations between objects and people in the visual environment, may actively influence incremental thematic role assignment and structural disambiguation.

Previous research has emphasized how interpretation of the linguistic input affects incremental or even anticipatory determination of reference to entities in the world during on-line comprehension (e.g., Altmann & Kamide, 1999; Chambers, Tanenhaus, Eberhard, Filip & Carlson, 2002; Hanna, Tanenhaus & Trueswell, 2003; Kaiser & Trueswell, 2002; Kamide, Altmann & Haywood, 2003; Kamide, Scheepers & Altmann, 2003; Runner, Sussman & Tanenhaus, 2003; Sedivy, Tanenhaus, Chambers & Carlson, 1999; Sussman & Sedivy, 2003). Additionally, evidence for the influence of visual referential contrast on comprehension processes comes from studies by Sedivy et al. (1999), and Tanenhaus et al. (1995). In reviewing these results, we identify three major strands of previous research.

Influential work by Tanenhaus et al. (1995) revealed the effects of referential visual contrast between two objects during on-line resolution of local structural ambiguity. They found that the initially adopted structure and interpretation is dependent upon the visual context. In instructions such as *Put the apple on the towel in the box*, the phrase *on the towel* can be temporarily interpreted as modifier of *apple* (identifying which apple) or as destination (where to put the apple). In a scene containing one apple, the phrase *on the towel* is preferentially interpreted as destination, as revealed by eye-movements to the towel-destination. In a scene contrasting two apples of which only one is on a towel, people preferentially interpret the phrase *on the towel* as modifier of *the apple*.

A second line of research has shown that visual context affects interpretation of temporary referential ambiguity in structurally unambiguous utterances. Sedivy et al. (1999) demonstrated that the time course of establishing reference to objects in a scene depended on whether there was referential contrast between the objects or not. For instructions such as *Pick up the tall glass and put it below the pitcher* in a scene with two tall objects, people looked at the target referent (the tall glass) more quickly when the visual context displayed a contrasting object of the same category (a small glass) than when it did not.

Both of these experiments provide evidence compatible with the Referential Theory of sentence processing (Altmann & Steedman, 1988; Crain & Steedman, 1985), and shed light on how reference to entities is determined incrementally. Furthermore, Sedivy et al. (1999) provide evidence for the view that referential processing is maximally incremental by demonstrating that listeners make use of referential contrast in the scene for establishing reference to an object prior to hearing the noun that refers to it.

A third line of work demonstrated that identification of relevant objects in a scene may sometimes "precede" their mention. For sentences containing a verb whose selectional restrictions applied to only one object out of several in the scene, such as *The boy will eat…,* Altmann and Kamide (1999) observed anticipatory eye-movements to the one appropriate edible object (a cake). No such anticipatory eye-movements to the cake were found when the verb's selection restrictions applied to several objects (a cake, a ball, a toy train, and a car) as was the case for sentences like *The boy will move….*

More recently, Kamide, Scheepers, and Altmann (2003) showed that unambiguous case-marking and verb selection restrictions influence expectations of post-verbal arguments and available referents. In German, a case-marked article can determine the grammatical function and thematic role of the noun phrase it modifies. Both SO (subject-object) and OS (object-subject) orders are grammatical. Participants inspected images showing a hare, a cabbage, a fox and a distractor object while hearing sentences such as *Der Hase frisst gleich den Kohl* ('The hare (subj) eats soon the cabbage (obj)') and *Den Hasen frisst gleich der Fuchs* ('The hare (obj) eats soon the fox (subj)'). Subject and object case-marking on the article of the first noun phrase together with verb selection restrictions allowed determination of the correct post-verbal referent. This was evidenced by anticipatory eye-movements to the cabbage after participants had heard 'The hare (subj) eats …' and to the fox after having encountered 'The hare (obj) eats …'.

In investigating diverse aspects of world-language relations, all of the above findings provide robust evidence for an account of incremental language processing in which processes of establishing reference to entities are central. While these are important, comprehension further entails the recovery of information such as grammatical function and thematic role, as well as propositional information. Previous research has, for example, revealed how compositional interpretation of the *linguistic input* triggers anticipation of an entity in the scene that fulfills thematic role requirements imposed by that linguistic input (Kamide, Altmann & Haywood, 2003; Kamide, Scheepers & Altmann, 2003).

There is little research, however, on the on-line influence of depicted events on comprehension processes such as incremental thematic role-assignment, and little is known about how reference to depicted events is established on-line. Indeed, world-language relations other than how language establishes reference to entities have received little attention in psycholinguistic research on on-line spoken sentence comprehension. One reason for this is that environments in preceding experiments in this research area only contained one type of visual information, namely entities (such as objects, animals, and characters). Clearly, a scene displaying only entities represents a highly restricted setting, severely limiting the influence that visual information can exercise on incremental comprehension.

Visual environments often offer a much richer variety of types of information, including actions and events. For the reminder of the paper, we concentrate on agent-action-patient events. Actions in such events determine a character's role (*carrying out an action* for agents), and establish role-relations between characters. If we see a woman greeting a man, the event information tells us that the woman is the agent of the greeting action, and the man the patient. Eye-movement studies of sentence production, for example, have already revealed a close link between visual attention in depicted action scenes and incremental production processes (e.g., Griffin & Bock, 2000). It should therefore be possible to understand role-relations established through actions when they are referred to by linguistic expressions such as verbs during comprehension. Making agent-action-patient events available in a scene might thus permit a more “active” influence of visual information on on-line comprehension processes, as they provide propositional as well as referential information.

Our experiments were designed to examine precisely whether visual role information provided by agent-action-patient events in the scene would guide incremental thematic role-assignment in the on-line comprehension of temporarily ambiguous utterances. If such depicted events influence incremental thematic role-assignment, then we argue that theories of sentence comprehension must incorporate richer semantic inventories than have been explicitly incorporated into theories of sentence processing to-date. In particular, they would need to be based upon a theory that allows reference to a variety of ontological categories. One framework, which fulfils this requirement, is Jackendoff’s semantic theory (1983, 1990, 2002). It allows us to describe reference to both actions and events and thus to account for how we access information about role-relations in the immediate visual context in a setting that offers combined linguistic and visual information. The studies reported in this article investigated the time course with which agent-action-patient events allow disambiguation of the local structural and role ambiguity that arises due to constituent order variability in German.

As noted above, German has a rich case marking system where grammatical function is usually indicated by case morphemes. Still, there is some case ambiguity. The subject case (NOM) and the object case (ACC) are identical for feminine noun phrases in German. While both SO and OS ordering is possible, SO is easier to process (Hemforth, 1993; Mecklinger, Schriefers, Steinhauer, & Friederici, 1995; Schriefers, Friederici & Kühn, 1995), and is the order used in the description of an event in isolated sentences. Case ambiguity and word order variation lead to grammatical function and role ambiguity.

In previous reading time studies on the SO/OS ambiguity in main clauses (e.g., Hemforth, 1993), participants usually interpreted a sentence-initial ambiguous noun phrase as subject, as evidenced by increased reading times when they encountered a second noun phrase that was marked as the subject. In German main clauses, the verb does not always disambiguate role and grammatical function of an ambiguous noun phrase that precedes it. People thus often realize only during a second noun phrase marked for subject-case that their subject-analysis of an ambiguous initial noun phrase was wrong. The experiments presented in this article investigate in particular if early disambiguation, prior to case-marking on the second noun phrase, was possible in the presence of disambiguating depicted events.

Experiment 1 examined the comprehension of initially structurally ambiguous spoken German SVO/OVS main clauses of the form NP1-V-ADV-NP2 (see Table 1). Sentences described event scenes showing a role-ambiguous character, acting and being acted upon (e.g., a princess), and two role-unambiguous characters that were agent and patient (e.g., a fencer and a pirate) respectively (see Fig. 1a). In contrast to previous studies investigating the influence of visual contrast between entities on how we establish reference to objects (e.g., Sedivy et al., 1999; Tanenhaus et al., 1995), reference to an entity was never ambiguous on the images in our experiments. Sentences and images only contained role-ambiguity. The first noun phrase of the sentence and its referent in the scene were always role-ambiguous, and the initial structural and role ambiguity in the linguistic stimuli persisted up to the second noun phrase, which was unambiguously case-marked as subject or object.

In contrast to studies by Kamide, Scheepers, and Altmann (2003), information about thematic role-relations between the sentential arguments was not provided by case-marking in the linguistic input prior to the second noun phrase. We ensured further that stereotypical relationships were absent from our materials. The relation between agents and actions (princess-washing), between agents and patients (princess-pirate), and between actions and patients (washing-pirate) were kept non-stereotypical. The absence of typical noun-verb relationships, and the presence of depicted actions that established thematic role-relations in the immediate visual context, crucially differentiates our experiments from previous studies by Altmann and Kamide (1999), Kamide, Altmann, and Haywood (2003), and Kamide, Scheepers, and Altmann (2003). In their experiments, the scene contained only objects, and thematic role-assignment was enabled through stereo-typical and selectional knowledge derived from the utterance. In our experiments, in contrast, thematic role-information was provided by the scene events alone. Early disambiguation and role-assignment would only be possible through the combined influence of sentence and the immediate visual context determining a character's role. The depicted actions differentiate between the event in which the princess is an agent (princess-washing-pirate) as described by SVO sentences, and the event in which she is a patient (fencer-painting-princess) as described by OVS sentences (see Fig. 1a, Table 1 (1a) and (1b)). Once the verb is encountered, the depicted actions that are necessary to establish role-relations between agents and patients in the events could be accessed.

If verb-mediated reference to depicted events enables early thematic role-assignment then expectations due to thematic role-assignment should be revealed in anticipatory eye-movements in the event scenes. When participants hear *Die Prinzessin wäscht* … ('The princess (amb.) washes …') for the SVO condition, with the princess-washing action depicting the pirate as patient, we expect more anticipatory eye-movements to the patient of the washing-event for SVO than for OVS sentences. In the OVS condition, the verb ('paints') does not match the action performed by the princess. Conversely, we expect a higher percentage of fixations on the agent for OVS than for SVO sentences after people heard *Die Prinzessin malt* … ('The princess (amb.) paints …'), with the fencer being identified as upcoming agent via the painting action (see Fig. 1a). Based on findings from Altmann and Kamide (1999), and Kamide, Scheepers, and Altmann (2003), eye-movements for early disambiguation should be anticipatory; that is, they should occur before participants heard the respective second noun phrase. We further expect this pattern to continue during the second noun phrase, which presents a further point of disambiguation through object or subject case-marking in the linguistic stimuli. Experiment 2 was aimed at replicating the findings from the first experiment while examining the influence of intonation. Experiment 3 examined processing of the active/passive ambiguity in German. Here sentences were also initially structurally and role ambiguous, but the main verb occurred in sentence-final position and hence did not establish reference to the depicted event information necessary for early visual disambiguation. In this case, we investigated whether a soft adverbial bias towards active or passive tense could lead to early structural and thematic role disambiguation before the second argument and main verb were encountered.

Experiment 1

Method

Participants

Twenty-eight German native speakers with normal or corrected-to-normal vision were paid five euro for taking part in the experiment.

Materials

We created 48 images using commercially available clipart and graphic programs. Images were pre-tested to ensure that participants were able to accurately recognize the agent-action-patient events and to discriminate between the two actions on an image. Sentences used in the pretest only differed from the experiment items in that they were all in SVO order. For each depicted action twenty participants saw both a matching and a non-matching sentence. For Figure 1a, participants read *Die Prinzessin wäscht/malt offensichtlich den Pirat* ('The princess (amb.) washes/paints apparently the pirate (ACC)') and *Der Fechter malt/wäscht offensichtlich die Prinzessin* ('The fencer (NOM) paints/washes apparently the princess (ACC)'). Participants were asked to judge if a sentence felicitously described an event on the image (*yes*) or not (*no*). The overall percentage of correct answers was 98.60.

A set of 24 items was created. Each item consisted of four spoken sentences and two pictures (see Table 1 and Fig. 1). One sentence per image had a canonical SVO and the other an OVS word order. Grammatical subject and object corresponded to agent and patient on the images respectively. Conditions were matched for length and frequency of lemmas using CELEX (Baayen, Pipenbrock & Gulikers, 1995) for all the words within an item. All the sentences had an SVO intonation up to the second noun phrase. For OVS sentences (e.g., Table 1. (1b)), the sentence beginning of the SVO sentence (e.g., Table 1, (2a)) was spliced in before the second noun phrase. This ensured that OVS intonation cues could not contribute to disambiguation.

There were two versions of each image, which only differed in the roles of the depicted characters; depiction of actions as such did not change. For example, one picture (Fig. 1a) showed a princess, a pirate and a fencer. The princess was washing the pirate, and the fencer was painting the princess. Thus, the princess was role-ambiguous (agent and patient); the pirate was the patient of the princess-washing action, and the princess was the patient of the fencer-painting action. For this scene, two sentences were recorded: *Die Prinzessin wäscht offensichtlich den Pirat*. ('The princess (amb.) is apparently washing the pirate (obj.)') and *Die Prinzessin malt offensichtlich der Fechter* ('The princess (amb.) is apparently painted by the fencer (subj.)') (see Table 1, (1a) and (1b) and Fig. 1a).

We expect comprehension processes of establishing reference and assigning thematic roles to manifest themselves in the eye-movement pattern when participants process sentences and scenes. When listeners hear 'The princess', we expect they will identify the princess as the appropriate referent in the scene. Hence, they should look predominantly at the princess. On hearing the verb 'washes' for the SVO sentence, it becomes clear that the princess is the agent of the washing event. As both, the orientation of the princess towards the pirate, and the depicted action provide information that the pirate is the target of the action, shortly after the verb, more eye-movements should be made to the pirate (the patient) than to the fencer (the agent). When the second noun phrase ('the pirate' (obj.)) is encountered, the fixations on the pirate should continue, since expectations of the pirate as the patient of the washing action will have been confirmed by object-case marking on the second noun phrase. For the OVS condition, people hear ‘The princess paints’. They presumably realize that the princess is not performing a painting action, and look for the agent of the painting event (the fencer). Shortly after the verb, participants should inspect the fencer more often than the pirate. On hearing ‘the fencer’ (subj.), their expectation of the fencer as the agent of the painting action is confirmed, and eye-movements to the fencer should further increase. Crucially, for SVO sentences, more anticipatory fixations should occur to the pirate as the patient of the event described by SVO sentences, than to the fencer prior to the second noun phrase. For OVS sentences, the fencer as the anticipated agent should receive more inspections than the pirate before the second noun phrase is encountered.

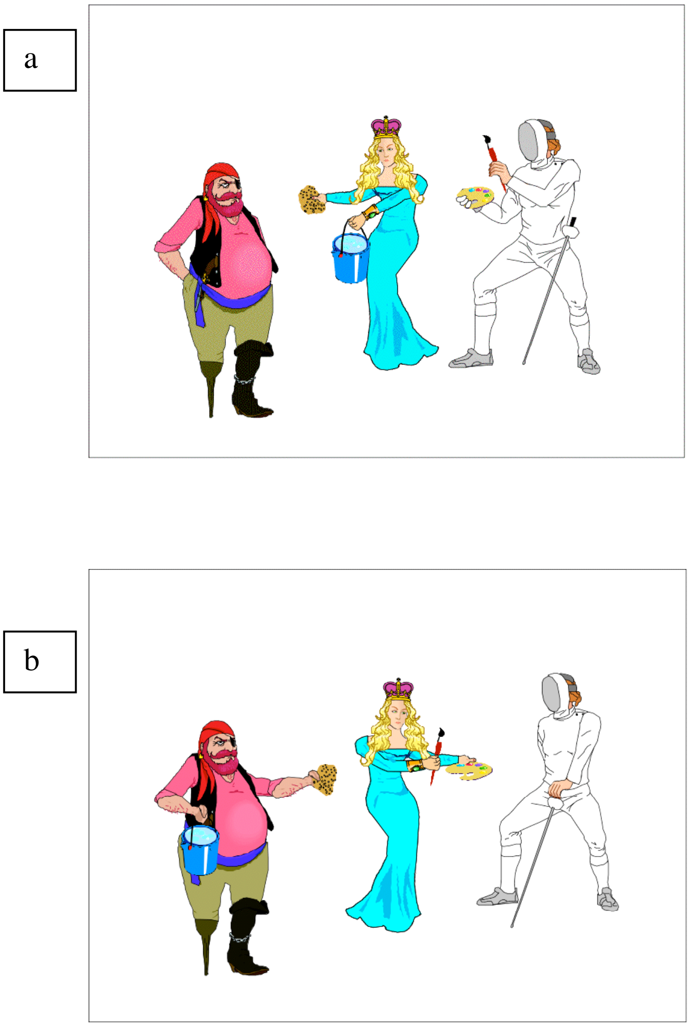


Figure 1: Example of picture pair for Experiment 1

Table 1: Example of sentence pairs for images in Figure 1

Image	Condition	Sentences
Figure 1: a	SVO	(1a) Die Prinzessin wäscht offensichtlich den Pirat. The princess (subject/object) washes apparently the pirate (object). 'The princess is apparently washing the pirate.'
Figure 1: a	OVS	(1b) Die Prinzessin malt offensichtlich der Fechter. The princess (subject/object) paints apparently the fencer (subject). 'The princess is apparently painted by the fencer.
Figure 1: b	SVO	(2a) Die Prinzessin malt offensichtlich den Fechter. The princess (subject/object) paints apparently the fencer (object). 'The princess is apparently painting the fencer.'
Figure 1: b	OVS	(2b) Die Prinzessin wäscht offensichtlich der Pirat. The princess (subject/object) washes apparently the pirate (subject). 'The princess is apparently washed by the pirate.'

In addition to the 24 items, a set of 32 filler items was constructed. Each filler item consisted of a scene and a sentence accompanying it. Eight started with an adverbial phrase and images showed two characters with only one being involved in an action; eight started with an unambiguously case-marked noun phrase and had four characters, with two being involved in an action; eight had scenes that did not contain action depictions, with verb-final sentences describing non-stereotypical actions in the future tense; eight had an unambiguously case-marked 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 the verb was not always in the second position; that there was not always an ambiguous noun phrase in the first position; that images did not always display a depicted action, and that there were not always three characters in the scene. Experimental items were separated from one another by at least one filler item. There were four lists of stimuli, each consisting of 24 experiment and 32 filler items. Each participant saw only one of the four conditions of each item, and the same number of items in each condition, and the order of items was randomized individually for every participant.

Procedure

An SMI EyeLink head-mounted eye-tracker with a sampling at a rate of 250 Hz monitored participants' eye-movements. Images were presented on a 21″ multi-scan color monitor at a resolution of 1024 x 768 pixels concurrently with the spoken sentences. Although only the dominant eye of each participant was tracked, viewing was binocular. Participants' head movements were unrestricted. Before the experiment, participants were instructed 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. They were shown two example images and sentences. Next, participants were set up and calibrated manually using a nine-point fixation stimulus. This procedure was repeated after approximately half of the trials. The EyeLink software validated calibration; if validation was poor, the calibration procedure was repeated until validation was good. Three filler items preceded the first experimental item for each participant. Between the individual trials, participants saw a centrally-located fixation dot on the screen, which they were asked to fixate. This allowed the eye-tracking software to perform a drift correction if necessary. The entire experiment lasted approximately 30 min.

Analysis

The procedure for analyzing the output data of the eye-tracker was as follows: The visual scenes were color-coded into distinct regions on bitmap templates (1024 x 786 pixels). The coded regions were the background and each of the three characters (e.g., for images in Fig. 1, the princess, the pirate, the fencer, and the background). The eye-tracker software recorded the *X-Y* co-ordinates of participants' fixations. This output was then converted into distinct codes for the characters and background so that participants' fixations were mapped onto the objects of an image.

Characters were coded depending on their event role for the inferential and descriptive analyses ('*amb. char., agent*, and *patient*, see Figs 2, 3, 4, and 5). Consecutive fixations within one object region (i.e., before a saccade to another region occurred) were added together, being counted as one *inspection*. Contiguous fixations of less than 80 ms were pooled and incorporated into larger fixations. Blinks and out-of-range fixations were added to previous fixations. We report the mean proportions of fixations on scene entities over the course of the entire utterance (Fig. 2), as well as inferential analyses and descriptive data of the proportion of inspections for individual time regions (Figs 3-5). We report proportions of inspections, a measure for which previous studies on auditory sentence comprehension in visual environments have established a close link to online comprehension processes (e.g., Tanenhaus et al., 1995; Sedivy et al., 1999).

For the descriptive presentation of the time course of the eye-movement data (Fig. 2), the process was a follows: For each time slot and object, the program computed the number of inspections that fell within a given time slot. For example, if an inspection on an object started at 1000 ms and lasted until 1125 ms after image onset, and time slots were 50 ms, then the program scored one inspection on that object for the 1000-1050 ms time slot, one inspection for the 1050-1100 ms time slot, and finally one inspection for the 1100-1150 ms time slot (i.e., the end of the inspection fell still within the 1100-1150 ms slot). For the subsequent slot from 1150-1200 ms the program would score zero inspections unless a new inspection started within that time slot.

Figure 2 plots the mean of the proportion of inspection counts 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.

In contrast to this, the data plotted for the individual time regions (Figs. 3-5) are based on exact trial-by-trial computations of these regions. Word onsets in each item speech file had been marked for the first noun phrase, the verb, the post-verbal adverb, and the second noun phrase. We computed the proportion of cases in a sentence condition for which there were inspections that started within a time-region (see Figs 3-5). The main time region for the analysis was the post-verbal adverbial region (*ADV*). As linguistic and visual information had to be integrated, we expected the disambiguation effect to occur shortly after the verb. The ADV region stretched from adverb onset to the onset of the second noun phrase (e.g., *apparently*, see Table 1) and had a mean duration of 794 ms for the SVO and 810 ms for the OVS condition.

The second noun phrase with its case-marking provided corroboratory evidence for disambiguation patterns on ADV and was thus included in the analysis (*NP2*). The NP2 interval extended from the onset of the case-marked article until 561 ms into the noun phrase. The mean duration of the second noun phrase was 818 ms. In addition we also examined inspections during the verb region (*VERB*), so as to check for potential early disambiguation effects on the verb (e.g., Altmann & Kamide, 1999). The mean duration of the verb was 491 ms for the SVO and 489 ms for the OVS condition. Temporal onsets of individual time regions within a trial were further calculated relative to image display onset by subtracting the image display onset from the respective time-region onsets.

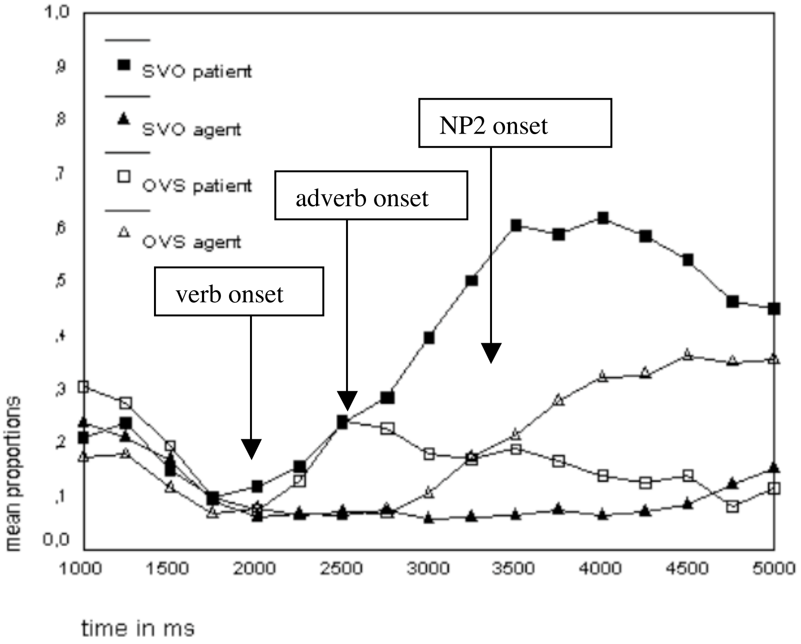
For the analysis of inspection proportions within a time region, we used hierarchical log-linear models, which combine characteristics of a standard cross-tabulation chi-square test with those of ANOVA. Log-linear models are adequate for count variables because they neither rely upon parametric assumptions concerning the dependent variable (e.g., homogeneity of variance), nor require linear independence of factor levels (Howell, 2001). Inspection counts for a time region were adjusted to factor combinations of target character (patient, agent), sentence condition (SVO, OVS) and either participants (N = 28) or items (N = 24). We report effects for the analysis with participants as LRχ^2(subj) and for the analysis including items as a factor as LRχ^2(item).

Results and Discussion

Figure 2 gives a detailed impression of the eye-movements over time by showing the mean proportion of inspections to characters during presentation of trials in time frames of 250 ms. The first graph of Figure 2 shows the eye-movements to the patient and agent on the images for SVO and OVS sentences from the onset of the first noun phrase (mean onset 1019 ms). The second graph displays inspections to the ambiguous character for both sentence types. Fixations on the background were left out for clarity of presentation, and inspections to the ambiguous character and the two target characters (agent, patient) are presented in separate graphs. The mean proportion of inspections to all entities (ambiguous, agent, patient, background) per sentence condition (SVO, OVS) add up to one hundred per cent.

The second graph of Figure 2 suggests that just before the onset of the verb, when participants have heard ‘The princess’, they predominantly inspect the ambiguous character (the princess). At the same time, they look equally often at the agent (the fencer) and patient (the pirate), but much less often than at the princess as shown in the first graph (Fig 2). The high amount of inspections to the ambiguous character immediately before and while hearing the verb (‘washes’/‘paints’) presumably simply reflects the fact that it has just been mentioned as first noun phrase. While people hear the verb, the pattern of eye-movements to the patient (the pirate) and agent (the fencer) changes. The first graph (Fig. 2) shows that shortly after the onset of the verb (‘washes’/’paints’), participants start to inspect the patient (the pirate) more often than at the agent (the fencer) for SVO and OVS sentences. This may be due to the interpretation of the sentence fragment, treating the first noun phrase as the agent and starting to look for the patient, but it may also be due to visual factors such as the gaze direction and posture of the ambiguous character towards the patient. Then, the patterns diverge at about the onset of the adverb. While participants who have heard SVO sentences (‘The princess washes’) continue to look at the patient of the washing-event (the pirate), participants who have heard OVS sentences (‘The princess paints’) start to look more at the agent of the painting event (the fencer). At the same time inspections to the patient (the pirate) decrease in the OVS condition (first graph of Fig. 2). Whereas the initial looks to the patient might be due to the orientation of the ambiguous character (the princess), this cannot be the case once the sentence type has an effect. At this point, we can be certain that the pattern of anticipatory looks reflects processes of identifying the verb with the appropriate depicted action.

During the second noun phrase, which provides another point of disambiguation at its case-marked article, the ambiguous character is looked at less often (see second graph of Fig. 2), as its mention occurred some time ago. The fixations on the agent have further increased when people heard ‘The princess paints apparently the fencer’ and exceed inspections to the patient (the pirate) for OVS sentences as shown in the first graph (Fig 2). The continuation of this fixation pattern during the second noun phrase provides clear evidence for disambiguation processes.



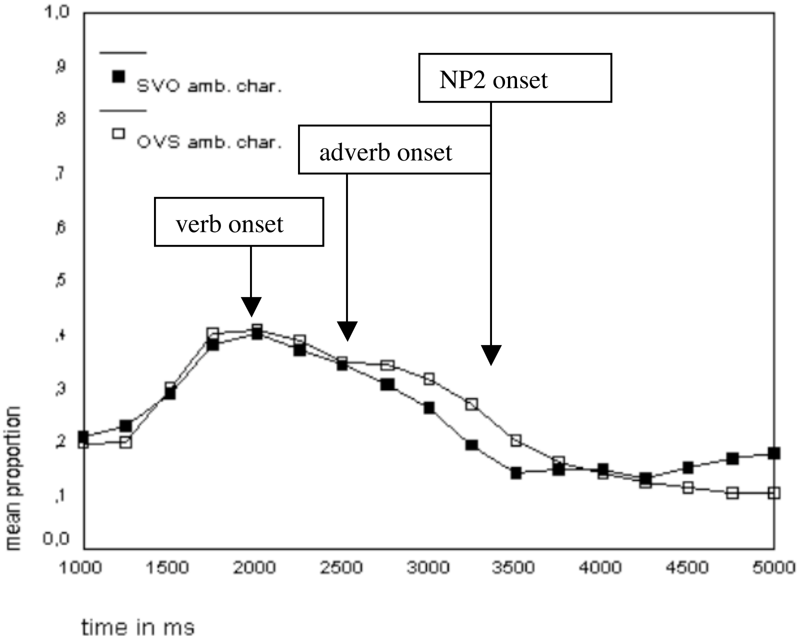


Figure 2. Time course of the eye-movement data for Experiment 1 showing the mean proportion of inspections to characters from the onset of the spoken stimuli in time frames of 250 ms

Participants are more likely to fixate the ambiguous character during the adverb time interval for OVS than SVO sentences. This coincides with the time that listeners are switching from fixating the patient (the pirate) to fixating the agent (the fencer). It may be due to participants who have not so far fixated the patient reprogramming their movement to fixate the agent, or it may be due to failed movements from patient to agent (participants reading SVO sentences are simply likely to continue inspecting the patient). Alternatively, it may be due to deliberate re-fixations on the ambiguous character in order to recompute its role from agent to patient. The main finding, however, is the evidence for tremendously rapid looks to the appropriate role filler for SVO and OVS sentences once the verb had been encountered.

We now turn to the inferential analysis. Crucially, the conclusions emerging from the descriptive analysis above hold for the inferential analysis of the data. While the time curves present an overview of the eye-movements to all scene entities, the inferential analyses for each experiment focus only on the theoretically most relevant effects. The proportion of inspections to all entities (ambiguous, agent, patient, background) per sentence condition (SVO, OVS) add up to one hundred per cent.

For the VERB region, log-linear analyses revealed a significant main effect of target character (patient, agent) (LRχ^2(subj) = 18.55, df = 1, p < 0.0001; LRχ^2(item) = 19.21, df = 1, p < 0.001). There was no interaction of sentence condition (SVO, OVS) and target character (agent, patient) (both LRχ^2s < 1). The main effect was due to a higher proportion of fixations on the patient than on the agent (see Fig. 3).

Figure 4 shows the proportion of inspections to the target characters and background during the ADV time interval for both sentence conditions (SVO, OVS). The interaction between sentence condition (SVO, OVS) and target character (agent, patient) was significant (LRχ^2(subj) = 82.30, df = 1, p < 0.0001; LRχ^2(item) = 83.81, df = 1, p < 0.0001). Log-linear contrasts confirmed that this was due to a significantly higher percentage of agent inspections in the OVS condition than in the SVO condition (LRχ^2(subj) = 39.95, df = 1, p < 0.0001; LRχ^2(item) = 40.23, df = 1, p < 0.0001), and to a significantly higher percentage of inspections to the patient in the SVO condition than in the OVS condition (LRχ^2(subj) = 68.15, df = 1, p < 0.0001; LRχ^2(item) = 67.86, df = 1, p < 0.0001). For the ADV time interval, contrasts showed in addition that there were a significantly higher percentage of inspections to the ambiguous character for the OVS condition than for the SVO condition (LRχ^2(subj) = 4.96, df = 1, p < 0.03; LRχ^2(item) = 4.95, df = 1, p < 0.03).

During the NP2 time interval, we found a significant interaction of sentence condition (SVO, OVS) and target character (agent, patient) (LRχ^2(subj) = 88.14, df = 1, p < 0.0001; LRχ^2(item) = 89.01, df = 1, p < 0.0001). This was due to a higher proportion of inspections to the patient for SVO than for OVS sentences (LRχ^2(subj) = 47.23, df = 1, p < 0.0001; LRχ^2(item) = 47.89, df = 1, p < 0.0001), and vice versa for the agent (LRχ^2(subj) = 75.94, df = 1, p < 0.0001; LRχ^2(item) = 77.37, df = 1, p < 0.0001) (see Fig. 5).

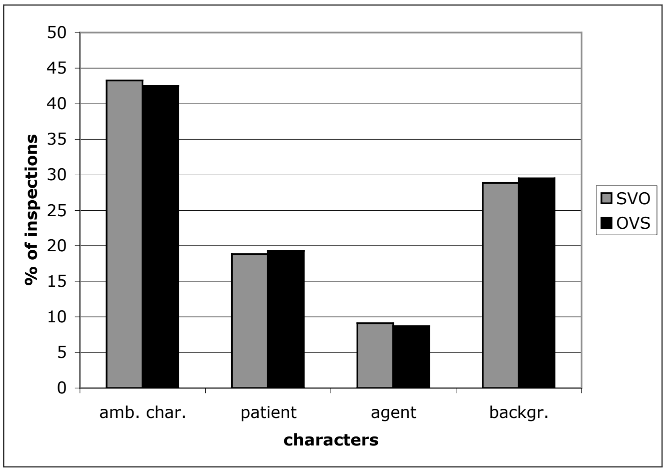


Figure 3: Percentage of inspections to characters for the VERB region in Experiment 1

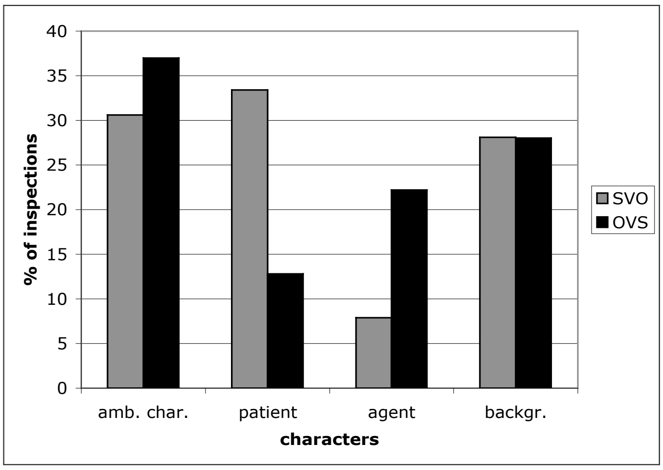


Figure 4: Percentage of inspections to characters for the ADV region in Experiment 1

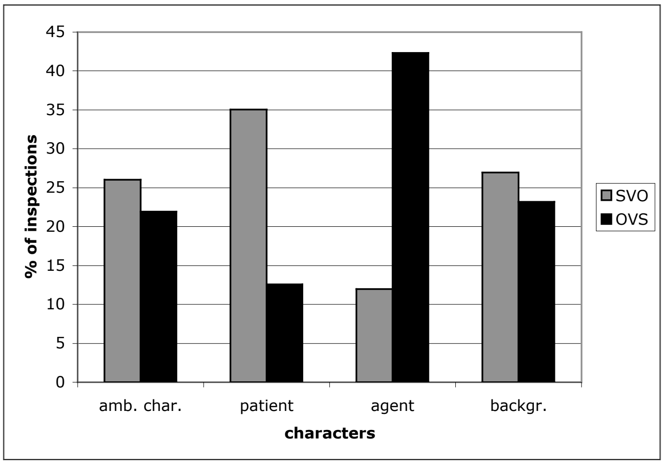


Figure 5: Percentage of inspections to characters for the NP2 region in Experiment 1

The fixation pattern prior to the second noun phrase could be due to either thematic role-assignment processes or a visual search strategy. A visual search account would argue that once people have heard the verb, they check whether it identifies the action performed by the first-mentioned entity (the princess). For SVO sentences, they hear 'washes', and this matches the princess-washing event, and hence they direct their attention to the character the princess faces (the pirate, Fig. 1a). For OVS sentences, they hear the verb 'paints', realize that the princess is not painting, start searching for an appropriate painting-action, and find the fencer-painting event. This account would moreover explain the rapidity with which people locate the appropriate depicted agent. It does not require that the fixations on the fencer for OVS sentences be guided by scene-derived expectations of thematic role-relations,

Alternatively, the observed eye-movements may reflect anticipation of thematic role-fillers, as argued in previous studies which did not explore the effects of depicted events but of case-marking and verb selection restrictions on thematic role-assignment (e.g., Kamide, Scheepers & Altmann, 2003). Such role-assignment account of our data would assume that people have extracted information about thematic-role relations from the event scenes prior to hearing the verb. In the SVO condition, when people hear 'The princess washes', they assign an agent role to the first noun phrase, as the princess is the agent of the verb ('wash'). They anticipate a patient role-filler and hence look predominantly at the patient of the washing action (the pirate) rather than the agent (the fencer). For OVS sentences, they hear 'The princess paints'. The assignment of an agent role to the first noun phrase is disconfirmed, as the princess is not doing the painting action. Previously acquired scene information about the fencer as the agent of a painting action, allows anticipating the fencer as the agent.

In order to dissociate between these two, possibly complementary, accounts, we consider research on visual and linguistic processing. Findings by Zelinksky and Murphy (2000) support the claim that linguistic properties of an object's name only influenced gaze behavior (frequency and duration of fixations) during object inspection when linguistic processing of objects was encouraged, not, however, when the task was visual search. Similarly, experiments on visual search and linguistic comprehension in text processing by Rayner & Raney (1996) found that word frequency only affected gaze duration when the task was reading, not during visual search. Accurate interpretation of gaze behavior during text and scene inspection hence necessitates consideration of the instructions given to participants. While the task in the present studies (scene and sentence comprehension) differs from those in the experiments by Zelinsky & Murphy (2000) (short term object recognition vs. visual search), an important commonality is that both scene/sentence comprehension and short-term object recognition encourage linguistic processing of the scene.

The fact that our instruction (understanding scene and sentences) encouraged linguistic processing thus argues for a thematic role-assignment rather than visual search account of our data. This line of reasoning is supported by results from a study that was carried out subsequent to the present experiments and with our materials (Scheepers & Crocker, in press). Their experiment replicated our findings of early visual disambiguation using inspection *duration* as measure, and further showed that the structure of a sentence presented prior to our stimuli had a priming influence on inspection duration for scene entities. Such priming presumably reflects linguistic expectations based on the preceding sentence structure, and further supports a contribution of thematic role-assignment in accounting for our data. Additional research is, however, necessary for a full understanding of how eye-movements in visual scenes depend upon the task at hand.

While a visual search strategy alone does not account for our data, it may nonetheless contribute to the observed gaze pattern. If so, this indicates a tight coupling of visual and linguistic processing, which ensures efficient scene perception and hence enables a most rapid impact of the perceived events on understanding. Expectation of thematic roles derived from scene inspection prior to hearing the verb may under this view combine with visual search as a rapid mechanism for locating scene objects that correspond to anticipated role-fillers.

Experiment 2

It is conceivable that some of the strength of our effects in Experiment 1 is due to a lack of distractor objects such as have been used in previous studies (e.g., Altmann & Kamide, 1999; Kamide, Scheepers & Altmann, 2003; Sedivy et al., 1999; Tanenhaus et al., 1995). After having encountered the ambiguous character, participants just had to choose between two possible options looking at the patient or the agent character), since there were no alternative objects to look at. Further, while partial cross-splicing in Experiment 1 excluded effects of the OVS intonation on disambiguation, we wanted to explore more fully the possible influence of intonation cues on the fixation patterns we observed by carrying out a fully cross splicing on the spoken materials.

Method

Participants

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

Materials

The design, procedure and analysis were the same as for Experiment 1. Two distractor objects were added on each image (see Fig. 6). Distractors were always inanimate, and were added to provide some further possibilities for people to look at. All filler items were modified to include distractor objects. While experimental sentences did not refer to the distractors, 12 of the 32 filler images had sentences that referred to them. Linguistic materials differed only in that we included the original version (unchanged intonation) and a cross-spliced version for each item sentence. For OVS sentences (e.g., Table 1, (1b)), the sentence beginning of the SVO sentence (e.g., Table 1, (2a)) was spliced in before the second noun phrase, and vice versa for SVO sentences. One item set thus consisted of eight sentences (four original, four cross-spliced) and two images. The cross-spliced versions for SVO and OVS sentences had an OVS and SVO intonation respectively up to the second noun phrase. The full cross-splicing allowed us to examine the influence of intonation cues on all the observed effects in a more complete way than in Experiment 1 where all sentence had SVO intonation up to the second noun phrase. Intonation (original, cross-spliced) was included as a factor in the inferential analyses.

The same word regions were chosen for analysis as in Experiment 1. The mean duration of the VERB interval was 490 ms for SVO, and 488 ms for OVS sentences. The mean duration of the ADV interval was 789 ms for SVO, and 798 ms for OVS sentences. The mean duration of the NP2 interval was 561 ms for SVO and OVS sentences.

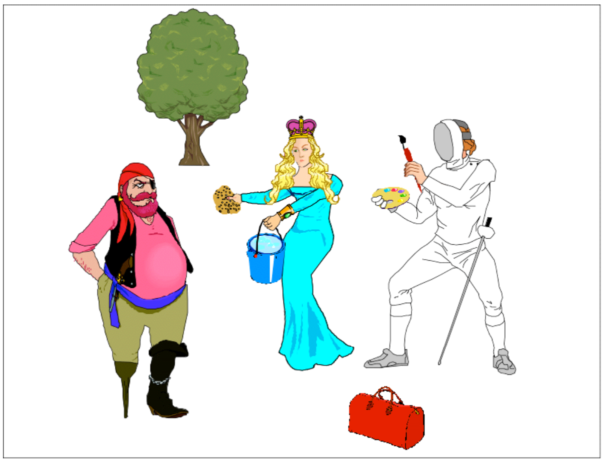


Figure 6: Example for image with distractor items

Results and Discussion

Figure 7 displays the time course of eye-movements during trials. The first graph shows inspections to patient/agent, and the second graph displays the time course of inspections to the ambiguous characters.

Overall, the picture that emerges is similar to the one for Experiment 1. Shortly after the onset of the verb (‘washes’/ ‘paints’), more inspections occur to the patient (the pirate) than to the agent (the fencer) as a result of either linguistic expectations or visual factors (see Table 1 (1a), (1b) and Fig. 1a). After the onset of the adverb, the fixation patterns diverge. For SVO sentences, participants continue to look at the patient, whereas in the OVS condition they start to look to the agent.

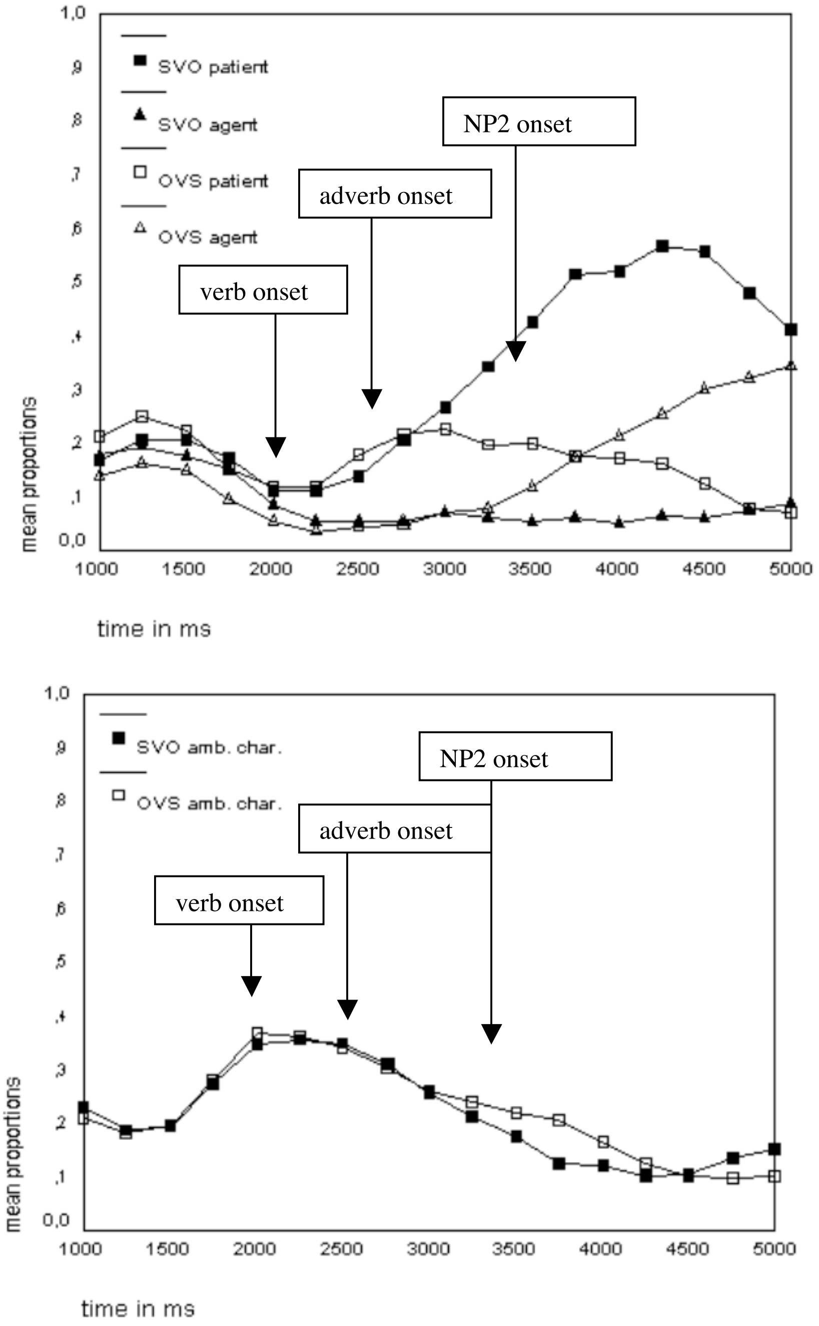


Figure 7. Time course of the eye-movements for Experiment 2 showing the mean proportion of inspections to each character from the onset of the spoken stimuli in time frames of 250 ms

The pattern of eye-movements during the second noun phrase after participants had heard the case-marked article was the same as for the adverb time interval, and provided further evidence for disambiguation processes.

We now turn to the inferential analyses. There was no effect of intonation (original, cross-spliced) for the analysis regions (all LRχ^2s < 1). The distractor objects were hardly looked at and had no visible influence on the eye-movement patterns during any of the time windows (see Figs 8, 9, and 10). For the VERB region, log-linear analyses revealed a significant main effect of target character (patient, agent) (LRχ^2(subj) = 27.07, df = 1, p < 0.0001; LRχ^2(item) = 29.46, df = 1, p < 0.0001) in the absence of a significant interaction (both LRχ^2s < 1). The main effect was due to a higher proportion of fixations on the patient than on the agent for both sentence conditions (Fig. 8).

Figure 9 shows the proportions of inspections to the characters, distractor objects, and the background during the ADV time interval. The interaction between percentage of inspections to characters (agent, patient) and sentence condition (SVO, OVS) was significant (LRχ^2(subj) = 106.95, df = 1, p < 0.0001; LRχ^2(item) = 111.31, df = 1, p < 0.0001) for this region. As for Experiment 1, log-linear contrasts revealed a significantly higher percentage of inspections to the agent for the OVS condition than for the SVO condition (LRχ^2(subj) = 69.12, df = 1, p < 0.0001; LRχ^2(item) = 69.69, df = 1, p < 0.0001), and vice versa for the patient in the SVO condition shortly after the verb (LRχ^2(subj) = 66.94, df = 1, p < 0.0001; LRχ^2(item) = 65.67, df = 1, p < 0.0001).

The elevated proportion of inspections to the ambiguous character for OVS as opposed to SVO sentences did not reach significance during the adverb time interval. It became significant during a later time window (from 200 ms before onset of the second noun phrase to 500 ms into the second noun phrase). It is unlikely that the delay was due to looks to the distractor objects, which are inspected very little during the adverb time interval (see Fig. 9). We assume that it is rather due to participants having to choose one out of five entities rather than one out of three as in Experiment 1.

During the NP2 time region, there was a significant interaction of sentence type (SVO, OVS) and target character (agent, patient) (LRχ^2(subj) = 129.32, df = 1, p < 0.0001; LRχ^2(item) = 136.92, df = 1, p < 0.0001). Contrasts again confirmed a higher proportion of looks to the patient for SVO than for OVS (LRχ^2(subj) = 93.72, df = 1, p < 0.0001; LRχ^2(item) = 96.97, df = 1, p < 0.0001) and vice versa for the agent (LRχ^2(subj) = 81.66, df = 1, p < 0.0001); LRχ^2(item) = 86.85, df = 1, p < 0.0001) (see Fig. 10).

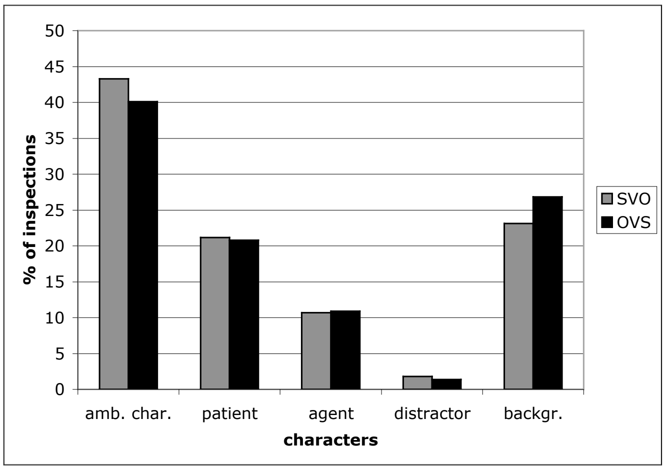


Figure 8. Percentage of inspections to characters for the VERB region in Experiment 2

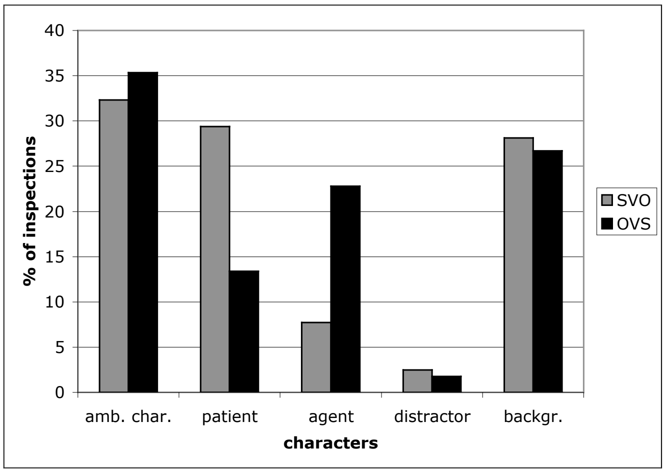


Figure 9. Percentage of inspections to characters for the ADV region in Experiment 2

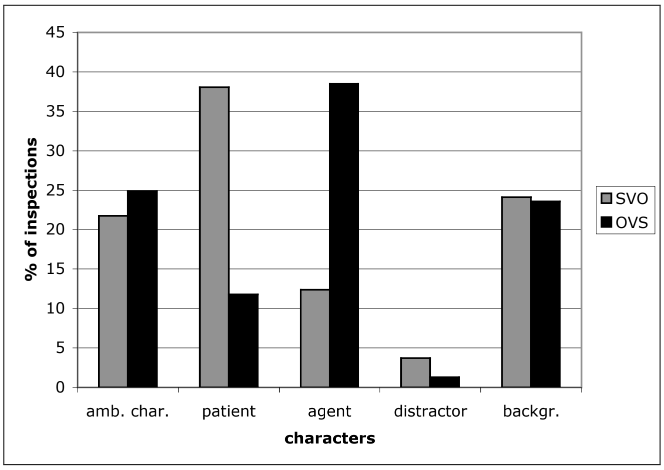


Figure 10. Percentage of inspections to characters for the NP2 region in Experiment 2

In sum, Experiments 1 and 2 provided strong evidence for early verb-mediated visual disambiguation of local structural and role ambiguity in German SVO/OVS sentences. Effects proved robust in the presence of distractor objects and a full cross-splicing.

Experiment 3

Experiment 3 investigated German verb-final active/passive sentences. As the verb occurred in sentence-final position, early reference to the depicted actions was not possible, and hence participants had to rely on purely linguistic cues for structural disambiguation. Two such cues were available in our experimental sentences. We expected linguistic disambiguation to occur when morphosyntactic marking on the second argument disambiguated the thematic role relations of the sentence. In order to explore possibilities of early linguistic disambiguation, we introduced a further soft adverbial constraint (temporal adverbs) prior to the second argument that biased towards either the active or passive event construction. In contrast to Experiments 1 and 2, where the ambiguity concerned grammatical function and role, resolving the active/passive ambiguity only involves assignment of the correct thematic role to the first noun phrase, and does not require changing its grammatical function. Observation of linguistic disambiguation in Experiment 3 would thus provide strong evidence that disambiguation patterns reflect thematic role-assignment processes.

Method

Participants

Twenty-six further participants from the same population as Experiment 1 were paid five euro for taking part in the experiment.

Materials

Images were the same as in Experiment 1. Spoken stimuli differed from those in Experiment 1 in two respects. We used a different type of construction (verb-final active and passive sentences) and introduced an additional adverbial manipulation. An example sentence pair for one image version (see Fig. 1a) is given in Table 2.

While German SVO/OVS sentences are ambiguous as to the grammatical function and role of the initial noun phrase, the ambiguity for the active/passive sentences in this experiment is purely a thematic role ambiguity. An initial subject noun phrase is disambiguated as agent in an active sentence when accusative (direct object) case-marking assigns a patient role to the second noun phrase. For passive sentences, preposition and dative case-marking disambiguate the thematic role of the second argument as agent, and the initial subject noun phrase receives the role of a patient (see Table 2). We use the term *second argument* to refer to both the noun phrase for active and the prepositional phrase for passive sentences.

The structural active/passive ambiguity relies on the auxiliary being ambiguous between a future tense auxiliary (*wird*, active ‘will’ or passive ‘will be’) and a present tense auxiliary (passive *wird*, ‘is’). Hence, if the auxiliary is interpreted as referring to the future tense, the initial noun phrase can be agent of an active or patient of a passive sentence; if it refers to the present tense it is the patient of a present tense passive sentence.

Table 2: Example sentences of one image version (see Fig. 1a) for Experiment 3

Image	Condition	Sentences
Figure 1: a	active	(1a) Die Prinzessin wird sogleich den Pirat waschen The princess (agent/patient) will soon the pirate (patient) wash. ‘The princess will soon wash the pirate.’
Figure 1: a	passive	(1b) Die Prinzessin wird soeben von dem Fechter gemalt. The princess (agent/patient) is currently by the fencer (agent) painted. ‘The princess is currently painted by the fencer.’

In order to explore other possibilities of early linguistic disambiguation, we introduced a further soft adverbial constraint. The hypothesis was that temporal adverbs bias interpretation of a preceding ambiguous auxiliary towards future or present tense, and hence influence choice of structure. We expected an adverb referring to a point in the future (e.g., *sogleich*, ‘soon’) to bias participants’ interpretation of the preceding *wird* towards an active future structure. While active or passive structures are possible continuations, people tend to opt for an active analysis in the absence of constraints biasing towards the passive. Present tense adverbs (e.g., *soeben* ‘currently’) should bias interpretation of the preceding auxiliary towards a passive auxiliary and hence towards a present tense passive structure. Adverbs were pre-tested for their bias and selected by means of sentence completions.

Twenty participants took part in the pretest. They read sentence beginnings such as *Sogleich wird die Prinzessin* … (‘The princess will soon…’) and *Soeben wird die Prinzessin* …(‘The princess is currently …’) on a computer screen. Participants were told to complete the sentence fragments. Sentences were then hand-coded according to whether completions were active or passive. We tested ten pairs of adverbs. For each pair, one adverb referred to the future, and one to the present. From these ten, we selected four future/present tense adverbial pairs that were matched for length and frequency (Frankfurter Rundschau Corpus, European Corpus Initiative, 1994). These were *sogleich/soeben* ('soon'/ 'currently'), *umgehend/momentan* ('immediately'/ 'currently'), *unverzüglich/im Augenblick* ('immediately'/ 'right now'), *sofort/derzeit* ('at once'/ 'at the moment'). The future tense adverbs used in the experiment resulted in 74% active completions, and 26% passive completions. The present tense adverbs resulted in 39 % active completions, and 61% passive completions. This was the strongest passive bias we were able to obtain under the additional constraints of length and frequency matching.

Based on results of Experiments 1 and 2, we predicted anticipatory inspections to the patient for both active and passive sentences on the auxiliary region and the ensuing adverb. We expected early linguistic disambiguation due to the adverbial bias before morphosyntactic marking on the second argument was available for disambiguation, if the manipulation was strong enough. Disambiguation of the initial role-ambiguity should definitely occur after participants heard the case-marked article/preposition of the second argument. We ensured that intonation did not influence disambiguation prior to the second argument. For the passive condition, all respective item sentences were recorded also with an active structure. The active sentence beginning was then spliced together with the passive continuations.

Procedure

The procedure and analysis were the same as for Experiment 1. The time interval (*AUXADV*) chosen for the analysis of the initial interpretation effect extended from the onset of the auxiliary (*wird* 'will'/'is') to the offset of the adverb (e.g., *sogleich* 'soon', *soeben* 'currently'). The rationale was similar to Experiments 1 and 2, in that anticipatory eye-movements to the patient were expected shortly after the first noun phrase. Given that the auxiliary region was very short, we included the adverbial region in the analysis. The mean duration of the AUXADV time region was 811 ms for the active and 788 ms for the passive condition.

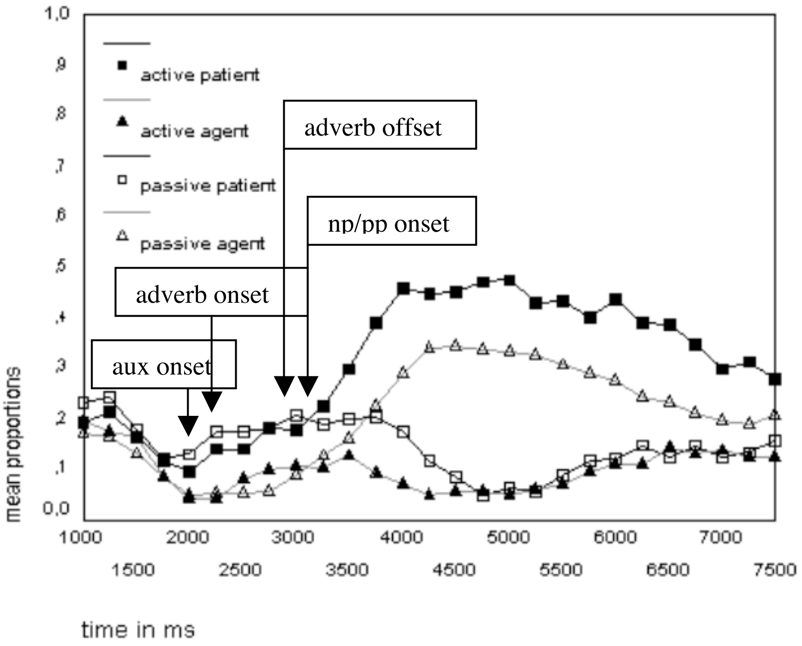
For early disambiguation by means of the temporal adverbs, we expected the effect to rapidly show up in the eye-movement pattern. The reason for this was that the adverbs did not refer to anything in the scene, and should provide a direct link to linguistic biases. We thus expected the effect to start once participants had heard most of the adverb. As inserting another filler word after the adverbs was not possible (it would have resulted in highly unnatural sentences), we ensured there was a short pause (mean duration of 200 ms) between adverb offset and onset of the second argument.

The time region for the inferential analyses corresponded thus to the late adverb (from 200 ms before the offset of the adverb for each item until the onset of the second argument) (*POSTADV*). Its mean duration was 401 ms for active and 422 ms for passive sentences. In order to ensure that early disambiguation (more looks to the agent as compared to patient) in the passive as opposed to the active condition could not be attributed to the additional 21 ms for the POSTADV region in passive sentences, we subtracted 21 ms from the onset of the second argument for every item in the passive condition. In the active condition the region extended up to the second argument for every item as described above. The POSTADV time region captured a time frame in which the adverbial constraints were available to the listener. At the same time, case-marking/preposition of the second argument was not yet available for disambiguation. As all sentences had active intonation up to the second argument, intonation cues could not influence disambiguation processes prior to the second argument. The region of the second argument (*SECOND ARGUMENT*) extended from the onset of the noun phrase/prepositional phrase to the onset of the sentence-final verb.

Results and Discussion

Figure 11 gives an impression of the time course of the eye-movements. The first graph displays looks to the agent and patient for both conditions, the second one displays inspections to the ambiguous character. The first graph shows that there are more inspections to the patient than to the agent for both active and passive condition during the AUXADV time interval (from a mean auxiliary onset of 2047 ms to a mean adverb offset at 2847 ms). The observed initial fixation pattern could, as in Experiments 1 and 2, be due to either visual factors or linguistic expectations of an active sentence construction as evidenced by past research (e.g., Kamide, Scheepers & Altmann, 2003; see also Ferreira, 2003; MacWhinney, Bates & Kliegl, 1984).

Between the offset of the adverb and the onset of the second argument (mean onset 3060 ms), the patterns begin to diverge. When hearing *Die Prinzessin wird gerade* ('The princess is currently'), participants start to look more at the agent for passive sentences, while looks to the patient start to fall.



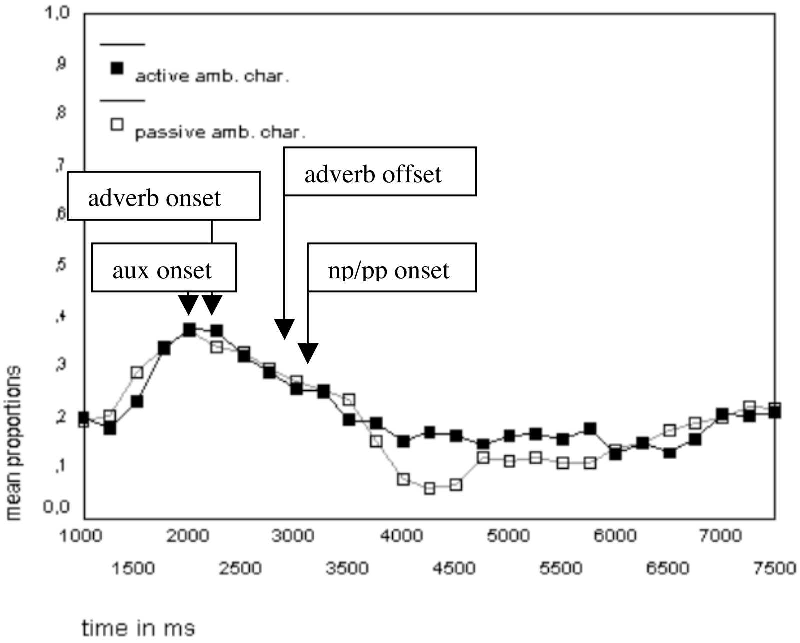


Fig. 11. Time course of the eye-movement data (mean proportion of inspections per time frame) to agent and patient for the active and passive condition presented in time frames of 250 ms for Experiment 3

They look more to the patient for active sentences when hearing *Die Prinzessin wird gleich* ('The princess will soon'), while simultaneously looks to the agent decrease. Disambiguation is clearly visible in the fixation patterns of the first graph during the second argument. The second graph of Figure 11 shows that participants looked more often to the ambiguous character for the active than for the passive condition around the onset of the second argument. This contrasts with findings from the first two experiments.

Log-linear analyses revealed a significant main effect of target character (agent, patient) on the AUXADV region (LRχ^2(subj) = 6.528, df = 1, p = 0.011; LRχ^2(item) = 6.604, df = 1, p = 0.010) in the absence of an interaction with sentence condition (active, passive) (both LRχ^2s < 1) (see Fig. 12). During the POSTADV time interval, the interaction between sentence condition (active, passive) and target character (agent, patient) was marginal (LRχ^2(subj) = 3.10, df = 1, p = 0.08; LRχ^2(item) = 3.50, df = 1, p = 0.06). In the log-linear contrasts only the analysis by items revealed a marginal effect of more looks to the patient for active than for passive sentences

($LR\chi^2$(item) = 2.71, df = 1, p = 0.05). For the other contrasts, effects were not significant (all ps > 0.1).

Interestingly, when only extending the POSTADV time region by 50 ms into the second argument (hence 50 ms into the second argument for active, and 29 ms into the second argument for passive sentences), effects become significant. There was a significant interaction between target character (agent, patient) and sentence condition (active, passive) ($LR\chi^2$(subj) = 4.79, df = 1, $p < 0.05$; $LR\chi^2$(item) = 4.87, df = 1, $p < 0.05$). The contrasts revealed a significant effect of more looks to the patient for active than for passive sentences ($LR\chi^2$(subj) = 4.37, df = 1, $p < 0.05$; $LR\chi^2$(item) = 5.25, df = 1, $p < 0.05$), a non-significant result for the agent inspections by subjects, and a marginal effect by items ($LR\chi^2$(item) = 3.14, df = 1, p = 0.08) (see Fig. 13).

An explanation of our findings in terms of intonation cues can be excluded as all sentences had active sentence intonation up to the second argument. As it is further estimated that programming of an eye-movement precedes the actual launching of a saccade by 200 ms (e.g., Matin, Shao & Boff, 1993), we argue that fixations during the first 29 or 50 milliseconds of the second argument cannot be due to the preposition/case marked article of the second argument. Rather, it seems likely that they have been caused by the preceding temporal adverbs.

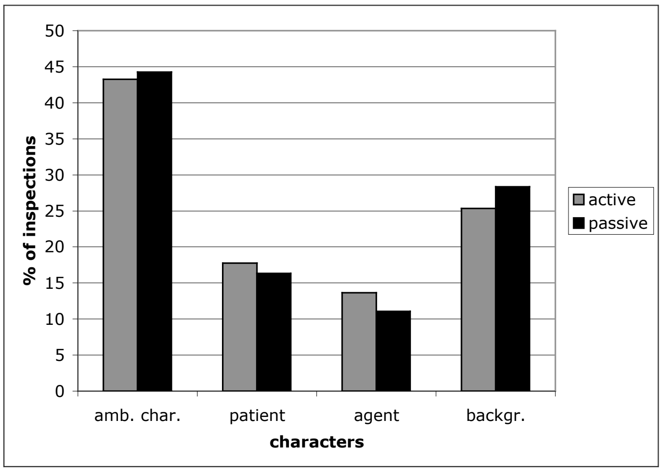


Figure 12: Percentage of inspections to characters for the AUXADV region in Experiment 3

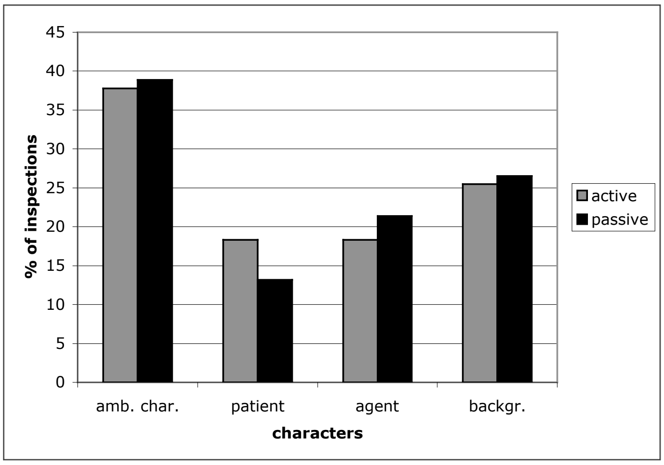


Figure 13: Percentage of inspections to characters for the POSTADV region in Experiment 3

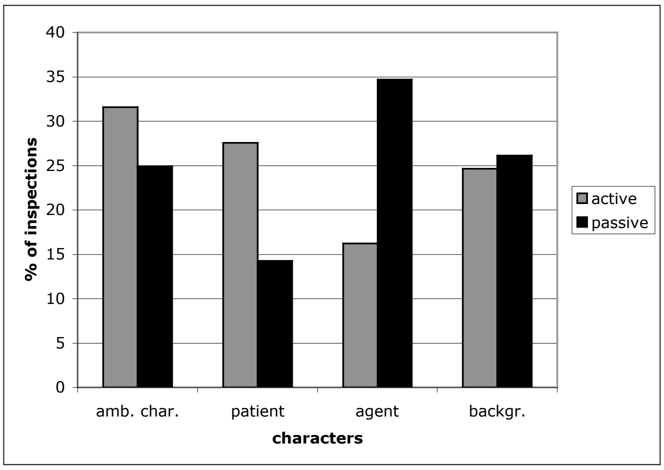


Figure 14: Percentage of inspections to characters for the SECOND ARGUMENT region in Experiment 3

The difference between the marginal log-linear contrasts for the patient inspections as compared to the non-significant result for the agent character might reflect the weakness of the passive bias versus the active bias. We assume that a difference of 20% between offline passive versus active completions for the passive bias was not large enough to trigger a significant effect in the contrasts. The interaction, however, provides evidence for early linguistic disambiguation.

During the SECOND ARGUMENT region, we found a significant interaction of sentence condition (active, passive) and target character (agent, patient) ($LR\chi^2$(subj) = 51.52, df = 1, p <

0.0001; LRχ^2(item) = 51.89, df = 1, $p < 0.0001$). Contrasts clearly confirmed more inspections to the patient for active than for passive sentences (LRχ^2(subj) = 27.11, df = 1, $p < 0.0001$; LRχ^2(item) = 27.23, df = 1, $p < 0.0001$), and a significantly higher proportion of fixations on the agent for the passive than for the active condition (LRχ^2(subj) = 42.83, df = 1, $p < 0.0001$; LRχ^2(item) = 43.17, df = 1, $p < 0.0001$). In addition, contrasts revealed significantly more looks to the ambiguous character for active than for passive sentences during this interval (LRχ^2(subj) = 5.71, df = 1, $p < 0.05$; LRχ^2(item) = 5.71, df = 1, $p < 0.05$) (see Fig. 14).

While participants heard the second argument, eye-movements further confirmed the disambiguation pattern. There were more looks to the patient than agent for the active as compared to the passive condition, and more looks to the agent than patient for passive as compared to active sentences (see Fig. 14). This pattern is the same as the one we observed for early visual disambiguation of the SVO/OVS ambiguity in Experiments 1 and 2. As only the thematic role and not the grammatical function of the first noun phrase is changed in disambiguation of the active/passive ambiguity, we can be certain that fixation patterns in Experiment 3 reflect thematic role assignment processes. There were, in addition, more looks to the ambiguous character for active than for passive constructions during the second argument, an effect that we did not observe for the first two experiments. We assume that it may be due to people anticipating the verb during the second argument, and hence looking at the ambiguous and agent characters' actions (as they were agent of the active and passive sentences respectively, and performing the actions).

General Discussion

Our findings show that perceived event relations between entities in the immediate visual context allow the rapid recovery of mental representations such as thematic roles in on-line comprehension of spoken sentences, and actively influence incremental resolution of initial structural and role ambiguity in the linguistic input.

This was revealed in Experiments 1 and 2, where anticipatory eye-movements in the event scenes provided evidence for expectations of a patient and agent role filler for initially ambiguous SVO and OVS sentences respectively once the verb had referred to the depicted actions. Experiment 2 replicated the early visual disambiguation effect in the presence of distractor objects and investigating more fully the possible influence of intonation. The rapid verb-mediated accessing of the depicted actions suggests further a tight coupling of visual and linguistic processing where a verb-initiated visual search cooperates with scene-derived expectations of thematic role-fillers in incremental interpretation of scene and sentence. Experiment 3, finally, demonstrated that when the verb was sentence-final and did not establish reference to the depicted events, linguistic cues alone enabled disambiguation. This was reflected by eye-movement patterns both during ambiguity resolution at the case-marked determiner/preposition of the second argument, and before this point as a result of the influence of temporal adverbs biasing towards the active or passive structure. This finding of early linguistic disambiguation from Experiment 3 further supports an account of our data in terms of thematic role-assignment, as the depicted event scenes were used for early thematic disambiguation before the verb had referred to the depicted actions.

In previous studies, the influence of visual contexts on the process of on-line ambiguity resolution came from contrast between entities when entities relied upon one differentiating attribute in determination of their referential uniqueness (Sedivy et al., 1999; Spivey, Tanenhaus, Eberhard & Sedivy, 2002; Tanenhaus et al., 1995). We suggest that such minimal contrast between object attributes (e.g., tall/small; on a towel/not on a towel) is one instance of how the perceived scene can establish (non-linguistic) relations between otherwise similar entities. Such contrastive relations arise through comparison by a perceiver, rather than being overtly realized in the scene, and hence represent an *implicit* relationship between entities. We propose agent-action-patient events are a further type, where perceived relations are *explicitly* realized visually by the depicted actions. Our results demonstrated how such explicit visual event relations influence structural analysis and interpretation of the linguistic input, and hence substantially extend the findings by Sedivy et al. (1999) and Tanenhaus et al. (1995). In building upon their work, we arrive at the view that non-linguistic information – such as contrast, actions, or events - that establishes relevant relations between entities, can affect how linguistic input is structured and interpreted. Emphasizing the importance of relations between entities makes it possible to account for both the effects of depicted event relations and visual referential contrast on ambiguity resolution.

Crucially, the presence of depicted agent-action-patient event relations provided role information about characters that could be accessed when the verb referred to a depicted action. The experiments we conducted hence extend previous findings in a further regard. Sedivy et al. (1999) provided robust evidence that when visual contexts contain referential ambiguity, referentially relevant visual information affects how reference to entities is determined incrementally. In our studies, images did not contain referential but rather role ambiguity, as one character on each image was both acting and being acted upon. The first two of our experiments therefore directly explored the as yet uncharted territory of how perceived event relations between entities in the visual context influence resolution of constituent-order, grammatical function and thematic role ambiguity on-line. They demonstrated that visually established role information determines incremental thematic role-assignment just as referentially relevant visual information has been found to influence processes of determining reference to entities in visual contexts.

In interpreting the anticipatory eye-movements to available role fillers in Experiments 1 and 2, we draw on research by Altmann and Kamide (1999) and Kamide, Scheepers, and Altmann (2003). Their experiments show that anticipatory eye-movements to role fillers in the scene were influenced by experience-based knowledge about the arguments a verb typically requires that could be extracted at the verb (e.g., Ferretti, McRae & Hatherell, 2001). Furthermore, investigation of the English active/passive ambiguity by Kamide et al. (2003) showed that tense marking disambiguates an initially role-ambiguous noun phrase (e.g., *The hare will eat…/The hare will be eaten…*). The cue for thematic role-assignment in their studies was the linguistic input. In contrast, for the first two experiments presented in this article, the linguistic input alone did not allow *early disambiguation*. Rather, once the verb had been processed, it was purely situation-specific visual role information in the event scenes, and verb-mediated identification of the appropriate event, that provided the information about a character's role necessary for early disambiguation.

There are theoretical consequences that follow from the observed influence of depicted actions on incremental interpretation. Previous studies (e.g., Sedivy et al., 1999; Tanenhaus et al., 1995) have drawn upon the Referential Theory of sentence processing (Altmann & Steedman, 1988; Crain & Steedman, 1985) as a framework for describing comprehension processes in visual contexts. Such a theory provides a suitable framework for the description of referential processes in sentence comprehension when reference is limited to the ontological category of entities. Our findings, however, suggest that on-line comprehension mechanisms exploit a wider range of ontological categories than just entities, including actions and events. In order to describe how depicted actions and events are exploited for on-line linguistic disambiguation, we need to establish reference to them through, for example, verbs and sentences. Theories of on-line sentence processing therefore require a suitably rich array of ontological categories and corresponding referential expressions, as well as mental representations for both.

Jackendoff's semantic theory allows us to straightforwardly include actions, events, and other ontological categories under the heading of reference. This follows from his assumption that the information conveyed by language is about the *perceptual world* (or *projected world*). He defines the perceptual world "not as absolute reality, but as the 'reality' constructed by our perceptual systems in response to whatever is 'really out there' " (2002, p. 308). He shows through both linguistic and non-linguistic arguments that reference in language must include a wide range of ontological categories among them places, actions, and events, and that "the class of referring expressions includes not only NPs but also Ss, VPs, PPs, and AdvPs" (1983, p. 41). On the premise that a semantic theory should explain otherwise arbitrary generalizations about the syntax and the lexicon, he maintains further that the array of mental representations "must contain a rich range of ontological categories" that correspond to different types of projected entities (1983, p. 56).

Support for the hypothesis that we structure the visual environment in the process of perception, and hence for Jackendoff's approach towards reference, comes from research on Gestalt Psychology (e.g., Köhler, 1947; Wertheimer, 1923/1938). Our findings are fully compatible with such a view. The insight from our results is that structuring of visual information in perception leads to the recovery of propositions rapidly enough to influence disambiguation and thematic role assignment processes on-line. Rather than seeing three unrelated characters (a princess, a pirate, and a fencer), two of whom are holding instruments, participants seem to have organized the image as displaying two events or three entities related to one another by actions. Ontological categories such as actions or events also play a key role in formal semantic theories such as Davidson (1967/1986a, 1969/1986b), or related approaches (e.g., Parsons, 1990; Carlson, 1984; Dowty, 1989). What makes Jackendoff's theory particularly suitable for our purposes, however, is that he situates language comprehension within a framework that includes other cognitive systems such as visual and auditory perception or the motor system. Furthermore, Jackendoff's notion of the *perceptual world* situates his approach in close proximity to an embodied view of the relation between language comprehension and visual environments (e.g., Bergen, Narayan & Feldman, 2003). His framework has the added advantage of offering an explicit and formal specification of an architecture and representations of the language faculty, enabling us to derive testable predictions in our current and future research.

We therefore adopt Jackendoff's architecture (1990, 2002) as a formal basis for describing the processing mechanisms that underlie on-line comprehension. Jackendoff (2002, p. 198) proposes a parallel constraint-based processing architecture, consisting of three independent but related levels of linguistic structure: phonological, syntactic, and semantic/conceptual structure. For each of these levels, there are constraints regarding processes internal to a level (*integrative*), and processes between the levels (*interface*). An *integrative* and an *interface* processor respectively accomplish internal processes and linking between the levels.

The individual levels in Jackendoff's architecture are modular in the sense of being domain-specific (i.e., their representational vocabulary is specialized), but unlike Fodorian modularity (Fodor, 1983), linguistic structures are linked among themselves and to other cognitive sub-systems. The version of modularity advocated in Jackendoff hence permits communication between phonological, syntactic, and conceptual structure. Importantly, it also allows information exchange between conceptual structure and perception or action via interface processors (2002, pp. 220f.). Jackendoff's framework provides an interface at which conceptual structure and visual information can be linked, and hence allows describing how comprehension proceeds when these different types of information are available and interact.

There are further requirements for the description of such interface processes. First, we need to explain how reference to various ontological categories in the perceived visual environment can be established. This is accomplished in Jackendoff (2002) via satisfaction of truth values. Just as nouns refer to conceptualized objects, sentences refer to conceptualized situations, with an additional layer that describes them as true or false depending on whether they are referentially satisfied or not. A partitioning of conceptual structure into tiers - descriptive (semantic/conceptual structure), and referential - allows linking of the individual tiers, and thus links referential information to linguistic structure. Second, we need to describe how visual information is encoded. Jackendoff (2002) proposes that visual information is encoded in *spatial structure*. Drawing on Marr's theory of vision (1982), Jackendoff (1987) describes in detail how the process of recovering visual information could proceed.

The processing architecture Jackendoff proposes is *logically non-directional* ((2002, p. 198), in principal permitting building of linguistic structure starting from vision or semantics and passing it on to syntax and phonology, or vice versa. The findings of our experiments are clearly compatible with this characteristic of his architecture. Both visual and linguistic information allowed disambiguation of initial structural and role ambiguity while visual scenes remained the same in all experiments.

Despite being logically non-directional, the parallel architecture put forward in Jackendoff (2002) does impose certain constraints on the sequential order with which the individual levels of linguistic structure and other cognitive systems can interact. Visual search singles out the correct interpretation of what we hear. Visual contexts cannot affect syntactic structure directly, but only through communication that is passed from the visual system via the visual-conceptual and the conceptual-syntactic interfaces to syntax. Inhibition of the incorrect alternatives is hence propagated to syntactic structure via the interfaces (2002, pp. 202f.).

In Experiments 1 and 2, it was purely verb-mediated depicted event information from the scene, which enabled structural disambiguation of the utterance. This finding clearly reveals an active influence of the immediate visual context on on-line role assignment and structural disambiguation. The insight that emerges from the third experiment is that such early influence of the depicted events on disambiguation does not, however, rely upon early verb-mediated reference to the relevant depicted event. When the verb occurred in sentence-final position, the depicted event relations could still be used for disambiguation, based on the active/passive bias occurring before the second argument and main verb. Data from Experiment 3 make it clear that in this case people relied on purely linguistic tense cues in the utterance in order to access the depicted event for resolution of structural and thematic role ambiguity. Taken together, our findings hence support a non-directional account of the interaction between visual and linguistic information in on-line sentence processing.

References

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., & 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). Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

Bergen, B., Narayan, S., & Feldman, J. (2003). Embodied verbal semantics: Evidence from an image-verb matching task. *Proceedings of the 25th Annual Conference of the Cognitive Science Conference*, 139-144.

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

Carlson, G. N. (1984). Thematic roles and their role in semantic interpretation. *Linguistics*, *22*, 259-279.

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

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.

Crystal, D. (2003). *A dictionary of linguistics and phonetics* (5th ed.). Oxford: Blackwell.

Davidson, D. (1986a). The logical form of action sentences. In: D. Davidson. *Essays on actions and events* (pp. 105-122). Oxford: Oxford University Press. (Reprinted from *The Logic of decisions and actions*, pp. 81-120, by N. Rescher, Ed., 1967, Pittsburgh: University of Pittsburgh Press)

Davidson, D. (1986b). The individuation of events. In: D. Davidson. *Essays on actions and events* (pp. 163-180). Oxford: Oxford University Press. (Reprinted from *Essays in honour of Carl G. Hempel*, pp. 216-234, by N. Rescher, Ed., 1969, Dordrecht: Reidel)

Dowty, D. (1989). On the semantic content of the notion 'thematic role'. In: G. Chierchia, B. Partee, & R. Turner (Eds.), *Properties, types and meanings* (pp. 69-129). Dordrecht: Kluwer.

European Corpus Initiative (1994). *The European Corpus Initiative Multilingual Corpus I* (CD-ROM). Available from Elsnet Web Site, http://www.elsnet.org/eci.html.

Ferreira, F. (2003). The misinterpretation of canonical sentences. *Cognitive Psychology*, *47*, 164-203.

Ferretti, T. R., McRae, K., & Hatherell, A. (2001). Integrating verbs, situation schemas, and thematic role concepts. *Journal of Memory and Language*, *44*, 516-547.

Fodor, J. (1983). *The modularity of mind.* Cambridge, MA: MIT Press.

Griffin, Z. M., & Bock, K. (2000). What the eyes say about speaking. *Psychological Science*, *11,* 274-279.

Hanna, J. E., Tanenhaus, M. K., & Trueswell, J. C. (2003). The effects of common ground and perspective on domains of referential interpretation. *Journal of Memory and Language*, *49*, 43-61.

Hemforth, B. (1993). *Kognitives Parsing: Repräsentation und Verarbeitung sprachlichen Wissens*. Berlin: Infix.

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

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

Jackendoff, R. (1990). *Semantic structures*. Cambridge, MA: MIT Press.

Jackendoff, R. (2002). *Foundations of language*. Oxford: Oxford University Press.

Kaiser, E., & Trueswell, J. C. (2002, March). *A new 'look' in the processing of non-canonical word orders: Anticipating upcoming referents on the basis of discourse status*. Paper presented at the 15th CUNY Conference on Human Sentence Processing, New York.

Kako, E., & Trueswell, J. C. (2000). Verb meanings, object affordances, and the incremental restriction of reference. *Proceedings of the 22nd Annual Conference of the Cognitive Science Society* (pp. 256-261). Hillsdale, NJ: Erlbaum.

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

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.

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

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

MacWhinney, B., Bates, E. & Kliegl, R. (1984). Cue validity and sentence interpretation in English, German, and Italian. *Journal of Verbal Learning and Verbal Behavior*, *23*, 127-150.

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

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.

Mecklinger, A., Schriefers, H., Steinhauer, K., & Friederici, A. D. (1995). Processing relative clauses varying on syntactic and semantic dimensions: An analysis with event-related potentials. *Memory & Cognition, 23,* 477-497.

Parsons, T. (1990). *Events in the grammar of English*. Cambridge, MA: MIT Press.

Rayner, K., & Raney, G. E. (1996). Eye-movement control in reading and visual search: Effects of word frequency. *Psychonomic Bulletin & Review, 3*, 245-248.

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(1)*, B1-B13.

Scheepers, C., & Crocker, M. W. (in press). Constituent order priming from reading to listening: A visual world study. In M. Carreiras & C. Clifton, Jr. (Eds.) *The online study of sentence comprehension: Eyetracking, ERP, and beyond*. Psychology Press.

Schriefers, H., Friederici, A. D., & Kühn, K. (1995). The processing of locally ambiguous relative clauses in German. *Journal of memory and Language*, *34*, 227-246.

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

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.

Sussman, R. S., & 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. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*, 1632-1634.

Wertheimer, M. (1938). Laws of organization in perceptual forms. In: W. Ellis (Ed.), *A source book of Gestalt psychology* (pp. 71-80). London: Routledge. (Reprinted from *Psychologische Forschung, 4,* pp. 301-350, 1923)

Zelinsky, G. J., & Murphy, G. L. (2000). Synchronizing visual and language processing: An effect of object name length on eye-movements. *Psychological Science, 11,* 125-131.

Acknowledgements

This research was funded by a PhD scholarship to the first author and by SFB 378 “ALPHA” to the second author, both awarded by the German research foundation (DFG).