Accessibility of Depicted Events Influences their Priority in Spoken Comprehension

Pia Knoeferle (knoeferle@coli.uni-sb.de)
Department of Computational Linguistics; FR 4.7, Building C71
Saarland University, Saarbrücken, Germany

Matthew W. Crocker (crocker@coli.uni-sb.de)
Department of Computational Linguistics; FR 4.7, Building C71
Saarland University, Saarbrücken, Germany

Abstract

Studies that monitor attention in depicted event scenes during utterance comprehension show that people prefer to rely on depicted events rather than their stereotypical knowledge. However, the presence of depicted events in the scene may have exaggerated their importance. Two eye-tracking experiments examined this issue by varying the accessibility of scene information. When we varied the visual presence of the scene (the scene disappeared before the utterance was heard), findings confirmed a greater relative priority of depicted events (Experiment 1). In contrast, when we altered the temporal extension of scene events (scene presentation emulated that they had been completed), people neither had a preference to rely on depicted events nor on their stereotypical knowledge (Experiment 2). These findings suggest that the visual presence of scene events cannot account for the preference to rely on depicted events. We discuss our findings in the context of research on the accessibility of events in discourse comprehension (e.g., Zwaan, Maaden, & Whitten, 2000).

The priority of depicted events

Prior research has revealed important insights into how we incrementally understand utterances that relate to a concurrently presented scene. In particular, the monitoring of eye movements in a scene while people listen to a concurrently presented and related utterance has shown that distinct types of information - linguistic and world knowledge as well as depicted events - are rapidly used for incremental thematic role assignment. This was revealed through the anticipation of relevant role fillers in the scene based on either linguistic and world knowledge (Kamide, Scheepers, & Altmann, 2003) or depicted agent-action-patient events (Knoeferle, Crocker, Scheepers, & Pickering, 2005) (see Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995, on the use of a visual referential context).

To further compare the *relative importance* of verb-mediated world knowledge about likely role-fillers with that of verb-mediated depicted events, Knoeferle and Crocker (accepted) carried out a study that directly compared the importance of these two informational sources. Participants in their study heard German object-verb-subject (OVS) sentences such as ‘The pilot (object) jinxes the wizard (subject)’ while inspecting the scene in Fig. 1. The noun phrase ‘the pilot’ established reference to the pilot in the scene. Object case marking on the determiner of the noun phrase together with the verb (‘jinx’) identified the pilot as the patient of a jinxing-action. The scene does not depict a jinxing action, but based on the verb ‘jinx’, we find a typical jinxing-agent (wizard) in the scene. In contrast, when we hear ‘The pilot (object) serves food to ...’, verb-based knowledge of stereotypical agents (e.g., a cook) cannot guide comprehension, as the scene affords no cook. However, the scene does afford a depicted food-serving event performed by the detective. Based on findings by Kamide et al. (2003) and Knoeferle et al. (2005), people should exploit either stereotypical knowledge of likely agents (‘jinx’) or depicted events (serving-food) for incremental thematic interpretation at the verb, and anticipate the appropriate agent.

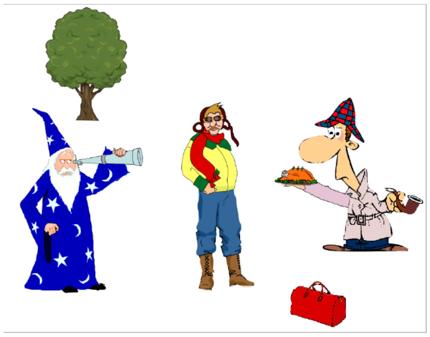

Figure 1: Example image for sentences in Table 1

In a further condition pair, people heard ‘The pilot (object) spies-on ...’. In this case, the scene affords both a stereotypical agent (the detective), and an immediately depicted agent of a spying event (the wizard) as relevant agents given the verb (see Fig 1). When people hear ‘spies-on’, word meaning, stereotypical knowledge of spy-on, and scene affordances allow them to anticipate either a depicted spying-event and its agent (the wizard), or a different, stereotypical agent (the detective). The comprehension system has to choose between two available, yet conflicting, types of information in online thematic role-assignment.

Gaze pattern confirmed prior findings that people rely on either stereotypical knowledge (Kamide et al., 2003) or depicted events (Knoeferle et al., 2005) when the verb (‘jinx’, ‘serves-food-to’) uniquely identified a relevant typical agent or the agent of a depicted (non-stereotypical) action. When the verb ('spy-on') identified, in contrast, two agents as relevant, there was a higher proportion of anticipatory eye movements to the agent depicted as performing the verb action (the wizard) in comparison with the agent that was stereotypical for the verb (the detective) shortly after the verb. This suggests a clear preference of the comprehension system to rapidly rely on depicted events over stored thematic knowledge for incremental thematic role assignment when both of these informational sources are available and identified as relevant by the utterance.

The origins of the preference?

Knoeferle and Crocker (accepted) suggest that the preference to rely on verb-mediated depicted events over verb-based stereotypical role knowledge has developmental origins. This position receives support from developmental research. Gillette, Gleitman, Gleitman, and Lederer (1999), while emphasizing the importance of linguistic knowledge for child language acquisition, propose 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 rock-bottom foundation for vocabulary acquisition" (Gillette et al., 1999, p. 169). Snow (1977) suggests that early language is largely about objects and events in the immediate environment of a child, thus corroborating this proposal (see, e.g., Yu, Ballard, & Aslin, 2005; Roy & Pentland, 2002, for relevant modeling research).

While such a developmental basis motivates the preference to rely on depicted events in adult comprehension, there are more subtle issues still to be resolved. On one account, the *visual co-presence* of a scene during comprehension may be the essential factor in the preference to rely on scene events over non-depicted stereotypical role relations between characters ("visual account"). The simultaneous presentation of a scene and a related utterance may have exaggerated the importance of depicted events. This is particularly true since - unlike child language acquisition - language in adult life often does not relate to an immediate scene.

Alternatively, what influences the preference to rely on depicted events may be the *discourse presence* of an event. In studies on discourse comprehension by Zwaan et al. (2000), participants read narrative texts that detailed a sequence of events. They had to decide whether or not an action word had occurred in either of two preceding sentences. Response latencies were longer when the prior discourse context indicated that an event had been completed, than when it was still ongoing. These results suggest that the temporal extension of an event (ongoing, terminated) influences its availability from the discourse context. In analogy, a scene event that has taken place, and is completed might be less available from the situation context ("temporal account").

The present paper examines whether it is the immediate visual presence of relevant depicted events, or the fact that they are depicted as ongoing in the situational context that accounts for the findings by Knoeferle and Crocker (accepted). To address this question, we conducted two eye-tracking experiments that varied the presence of scene information during comprehension.

Prior scene context

There have been first steps towards exploring how people access information about objects when those objects are not immediately present in the scene. In Spivey and Geng (2001, Exp. 2), scenes contained four differently coloured shapes that were either tilted leftward or rightwards. After participants had inspected the objects, one of them disappeared. When answering a question about either colour or tilt of the disappeared object, participants re-fixated its previous location on half of the trials. It has been proposed that people use pointers to the external world to store object properties (Richardson & Spivey, 2000; Spivey & Geng, 2001).

Altmann (2004) extended their approach to the investigation of utterance comprehension. People inspected an image with a man, a woman, a cake, and a newspaper, and then the screen went blank ("blank-screen" paradigm). Two seconds later, people heard a sentence that described part of the previously-inspected scene (e.g., *The man will eat the cake*). Having heard the verb, people rapidly looked at the location where previously there had been a cake. These data suggest that even when a prior scene context is no longer immediately available, the mental representations that listeners retrieved from it may still inform a listener's expectations about objects that are likely to be mentioned next.

While these findings make it clear that information about previously inspected objects informs utterance comprehension, the *extent* to which scene information is available for comprehension when it is not immediately visually present is unclear. It is further unclear whether findings by Altmann (2004) extend to more complex scenes that contain agent-action-patient events in addition to objects and characters.

Experiment 1 was designed see whether findings by Altmann (2004) extend to richer scenes, and to test the visual account of the preference to rely on depicted events. People inspected a scene (Fig. 1), and then the entire scene disappeared prior to utterance presentation. If the visual account were most appropriate, we would not expect to replicate the greater relative priority of depicted events in this setting.

Experiment 1

Method

Participants Thirty-two native speakers of German with normal or corrected-to-normal vision received five euro each for participating in the experiment.

Materials and Design The materials were the same as in Knoeferle and Crocker (accepted). There were 24 items and four experimental conditions (Table 1, (a1), (a2), (b1), (b2), and Fig. 1). We describe the four experimental conditions, and then the counter-balancing.

An example image showed two agents (e.g., a wizard, and a detective), each performing an action upon a patient (e.g., the pilot, Fig. 1). The depicted events provided information about role relations (e.g., wizard-spying-on-pilot and detective-serving-food-to-pilot, see Fig. 1). In addition, each agent provided stereotypical thematic role knowledge (e.g., a detective is a stereotypical agent of a spying, and a wizard a stereotypical agent of a jinxing action).

All sentences were unambiguous OVS sentences. They always started with an unambiguously accusative case-marked noun phrase referring to a patient role-filler (Fig. 1, the pilot). While the OVS structure is non-canonical in German, this was a constant across conditions.

Image	Cond.	Sentences
Fig.1	(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.1	(a2)	Den Piloten verzaubert gleich der Zauberer. The pilot (PAT.) jinxes soon the wizard. The wizard will soon jinx the pilot.
Fig.1	(b1)	Den Piloten bespitzelt gleich der Zauberer. The pilot (PAT.) spies-on soon the wizard. The wizard will soon spy on the pilot.
Fig.1	(b2)	Den Piloten bespitzelt gleich der Detektiv. The pilot (PAT.) spies-on soon the detective. The detective will soon spy on the pilot.

Table 1: Example item sentence set for Fig. 1

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

Materials were designed so that plausibility biases introduced by the verb or noun phrases, depiction biases introduced by characters, or position biases (whether an agent was to the left or right of the patient) were counter-balanced. Balancing further ensured that each target agent (wizard, detective) was once identified as a potential agent through verb-mediated depicted actions and through stereotypical verb-based knowledge. Each agent was once uniquely identified as relevant by the utterance, and once identification of the relevant agent was ambiguous. Conditions were matched for length (number of spoken syllables), and frequency of lemmas within an item (t-test, p = 1) (Baayen, Pipenbrock, & Gulikers, 1995). Off-line plausibility ratings ensured that the character which was the stereotypical agent for a verb (detective, *spy-on*) was a more plausible agent for that verb than the other character (wizard, Fig. 1).

There were 48 filler items, balanced for the number of stereotypical versus depicted actions and agents. Twelve trials showed an agent performing a stereotypical action to avoid that participants pay less attention to their stored linguistic and world knowledge than to the (implausible) depicted events. To ensure that the majority of sentences a participant heard were canonical, forty filler utterances were subject-initial sentences. There were eight experiment lists. Consecutive experiment trials were separated by at least one filler trial. Each of the 32 participants saw only one of the four conditions of an item, and the order of appearance of items was randomized individually for every participant.

Procedure An SMI Eye-Link head-mounted eye-tracker monitored participants eye-movements at a frequency of 250 Hz. Images were presented on a 21-inch multi-scan color monitor at a resolution of 1024 x 768 pixel. Before the experiment, participants received written instructions. They were asked to inspect the images attentively, and informed that the study examined what happened when people listen to sentences that describe events that either are depicted on a previously viewed scene, or that could take place. For each trial, the image was presented first, and participants inspected it for a fixed duration of 7000 ms. The image then disappeared, and the screen went blank. After 1500 ms, the utterance was presented (see Altmann, 2004, for similar instructions and a similar delay). The entire experiment lasted approximately 30 min.

Analysis For the analyses we chose a time region for which participants had heard most of the verb, but not the second noun that identified a depicted or stereotypical target agent. This region extended from 250 ms before the offset of the verb to the offset of the adverb (ADV) for each item trial (see Knoeferle & Crocker, accepted). The X-Y coordinates of participants fixations were assigned to regions that corresponded to entities and scene background of the previously inspected scene. For Fig. 1 the wizard was coded as stereotypical agent, and the detective as depicted agent for the unique conditions ((a1) and (a2) respectively), and vice versa for the ambiguous conditions ((b1) and (b2), see Table 1). Consecutive fixations within an object region (i.e., before a saccade to another region) were added together as one *inspection*. For the inferential analysis we used hierarchical log-linear models (see Howell, 2002; Knoeferle et al., 2005). Inspection counts for the ADV region in the unique (Table 1, (a1), (a2)) and ambiguous conditions (Table 1, (b1), (b2)) were subjected to analyses with the factors *target character* (stereotypical agent, depicted agent), *target type* (depicted, stereotypical), and either *participants* (N = 32) or *items* (N = 24).

Results and discussion

Figs 2 and 3 show the proportion of inspections to the target characters (depicted agent, stereotypical agent) during the ADV region. Fig. 2 shows inspection percentages to the entities for the unique ((a1), (a2), Table 1), and Fig. 3 for the ambiguous conditions ((b1), (b2)).

For the unique conditions a significant interaction of target type (depicted, stereotypical target) and target character (depicted agent, stereotypical agent) revealed that people can rely on either depicted events (a1, Table 1) or stereotypical knowledge (a2) in anticipating a relevant agent ($ps < 0.001$). There was a higher percentage of inspections to the stereotypical agent in the stereotypical (a2) than in the depicted target condition (a1), and more inspections to the depicted agent in the depicted target (a1) than in the stereotypical target condition (a2) (Fig. 2, Table 1).

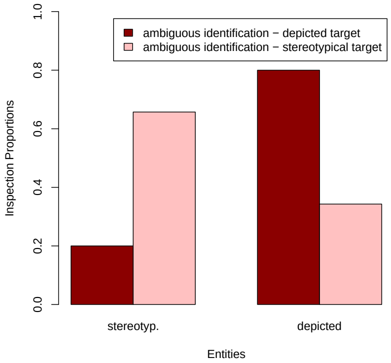

Figure 2: Gaze proportions during ADV for the unique conditions in Experiment 1

For the ambiguous conditions (Table 1 (b1), (b2)), we found - as expected - no interaction for the ADV region since sentences were identical prior to the second noun. Rather, we replicated the main effect of significantly more inspections to the depicted than to the stereotypical agent for both (b1) and (b2) (Fig. 3). These looks occurred after people had heard 'The pilot (object/patient) spies-on ...' and before they heard the respective second noun phrase, which then disambiguated towards the depicted or stereotypical target type. Log-linear analyses showed that the main effect of depicted agent was significant ($p < 0.001$ by part., $p = 0.08$ by items) in the absence of a significant interaction ($ps > 0.4$).

The main effect for the ambiguous conditions (b1 and b2) is only meaningful in comparison with the significant interaction found in the unique conditions (a1 and a2). A three-way interaction between target character (depicted agent, stereotypical agent), identification (unique, ambiguous), and target type (stereotypical, depicted) confirmed that the difference between the main effect in the ambiguous an the interaction in the unique conditions was significant ($p < 0.05$ by part., $p < 0.01$ by items).

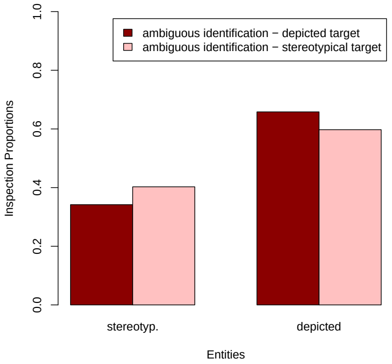

Figure 3: Gaze proportions during ADV for the ambiguous conditions in Experiment 1

The above analyses confirmed findings of a greater relative importance of (non-stereotypical) depicted events over stereotypical knowledge of (non-depicted) events during online comprehension (see Knoeferle & Crocker, accepted). This suggests that the immediate visual presence of the scene is not a necessary pre-requisite for the greater relative importance of depicted events.

Experiment 2

Experiment 2 tested the alternative, temporal account. Scene presentation was modified so as to emulate that the depicted events had been completed, hence making them less available in the situation context. For Experiment 2, people inspected a scene like Fig. 1 but without the actions. Then we briefly depicted the actions, removed them, and only the characters (and their stereotypical affordances) remained on the display during utterance presentation. In this manner, we emulated that an event had taken place and was completed. If people still prefer to rely on depicted events this would be evidence against both a visual and temporal account. Finding, in contrast, that the priority of depicted events is diminished in this setting, would suggest that the perceived temporal extension of a scene event in the context (ongoing, completed) influences its accessibility and - in our study - priority for spoken comprehension.

Method

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

Materials, Design, and Procedure Materials and design were identical to Experiment 1. The only difference between Experiments 1 and 2 concerned image presentation. In Experiment 1 the entire scene disappeared after an inspection time of 7000 ms. In Experiment 2, scene presentation consisted of four steps. First, participants inspected the scene in Fig. 4a. The scene characters were identical to the ones on the image for Experiment 1 (Fig. 1); however, unlike Fig. 1, the initial scene in Experiment 2 did not show depicted actions. After 1500 ms, one action appeared (wizard, spying with a telescope, Fig. 4b). After another 1500 ms, the action of the other agent (detective, serving food on a plate, Fig. 4c) was presented. Finally, both actions disappeared together, and only the characters remained in the scene during utterance presentation (Fig. 4d).

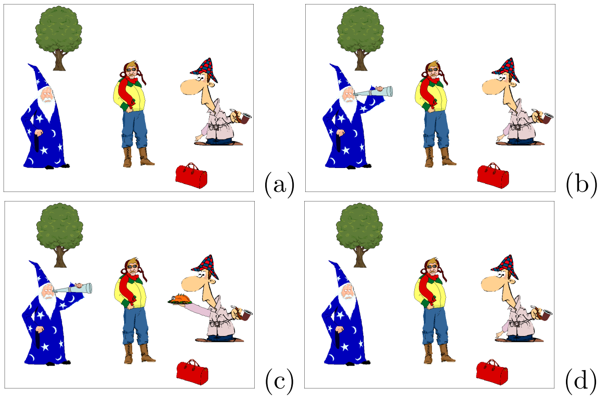

Figure 4: Sequence of four scenes in Experiment 2

For the first three scenes (Figs 4a-c) no speech input was presented. To balance directionality effects, the action appeared first for the rightmost agent, and subsequently for the leftmost agent for one half of the items; for the other half, the action appeared first for the agent to the left, and then for the agent to the right of the patient. The filler images and sentences were the same as for Experiment 1. Presentation was, however, changed to resemble that of the experimental images. Thus, for each filler trial, participants saw a sequence of four related scenes on which objects and characters appeared, moved, or disappeared. For analyses we used the same ADV region as for Experiment 1.

Results and discussion

Figs 5 and 6 show the proportion of inspections to the target characters (depicted agent, stereotypical agent) during the ADV region. Fig. 5 shows inspection percentages for the unique ((a1), (a2), Table 1), and Fig. 6 for the ambiguous conditions ((b1), (b2)).

For the unique conditions, log-linear analyses revealed a significant interaction of target type (depicted, stereotypical target) and target character (depicted agent, stereotypical agent) ((a1) and (a2), Table 1) ($ps < 0.05$). The interaction resulted from a higher percentage of inspections to the stereotypical agent for the stereotypical (a2) than the depicted target condition (a1), and an increased proportion of inspections to the depicted agent for the depicted (a2) than the stereotypical target condition (a1) (Fig. 5).

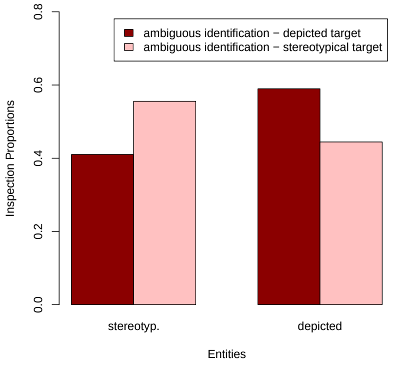

Figure 5: Gaze proportions during the ADV for the unique conditions

In the ambiguous conditions ((b1), (b2)), unlike for Experiment 1, we find no main effect of target agent ($ps > 0.14$). There was no reliable interaction between target type (depicted, stereotypial) and target agent (depicted agent, stereotypical agent) ($ps > 0.2$) (Fig. 6).

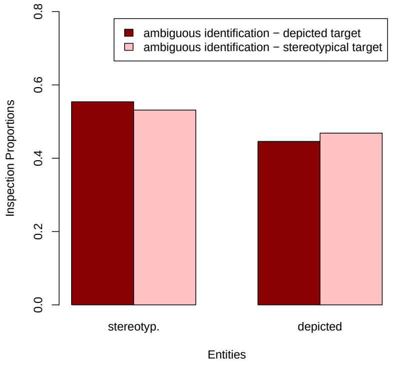

Figure 6: Gaze proportions during the ADV for the ambiguous conditions

The analyses support the view that when scene presentation emulated that depicted events had been completed, the preference to rely on depicted events for thematic interpretation in the ambiguous conditions was reduced. This provides evidence for the temporal account.

General Discussion

Our findings support the view that it is the accessibility of an event in the situational context (i.e., whether it is depicted as ongoing or presented as completed) that is at the origin of the preferred reliance on verb-mediated depicted events over verb-based stereotypical knowledge of events.

Experiment 1 found that the visual presence of the scene is not a pre-requisite for the preferred reliance on depicted events. People anticipated either the agent of a non-depicted, stereotypical action ('jinx', wizard) or the agent of a depicted action ('serves-food-to', detective) when the verb uniquely identified an agent as relevant. In contrast, when the verb ('spies-on') identified two different agents as relevant, people anticipated the agent of the depicted action (wizard) more often than the stereotypical agent (detective). Gaze pattern showed that the findings by Knoeferle and Crocker (accepted) extend to the blank-screen paradigm (see Altmann, 2004).

Findings from Experiment 2 provide more detailed insights into the use of depicted events. When events depicted were presented in a manner that emulated events taking place and then being completed, and characters (and their stereotypical affordances) remained present during utterance presentation, people were still able to use verb-based stereotypical thematic knowledge or depicted events when the verb uniquely identified either a stereotypical agent or the agent of a depicted action as relevant for thematic interpretation. However, when the verb ('spies-on') identified two different agents as relevant, we found no preferential anticipation of either a stereotypical or a depicted agent (see Fig. 1).

Together, gaze patterns from Experiments 1 and 2 suggest that visual availability cannot account for findings by Knoeferle and Crocker (accepted). If a visual account was correct, we should not have replicated the existing findings in Experiment 1, and we would have expected to find a preferred reliance on the more immediately available stereotypical role fillers in the ambiguous conditions of Experiment 2. Rather, the findings are compatible with a temporal account that relies on whether a scene event is ongoing or completed in the situation context.

Parallels to discourse contexts Our research on using a prior scene context for utterance comprehension builds on existing work that has examined how people use information about objects and events from prior discourse. Findings by Zwaan et al. (2000) suggest that the temporal extension of an event (ongoing, terminated) influences its degree of availability from the discourse context (see also, e.g., Magliano & Schleich, 2000). This insight is relevant for our findings since scene presentation in Experiment 2 (Fig. 4) was manipulated to create an impression of events being completed. This completed status may have diminished the accessibility of the events, in particular also since the verb and adverb in our sentences referred to a future action (e.g., 'spies-on soon'). Future studies will vary verb tense to explore this possibility.

Acknowledgements

This research was funded by a PhD scholarship to PK, and by SFB 378 ALPHA to MWC, both awarded by the German Research Foundation.

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.

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

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

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

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., & Crocker, M. W. (accepted). The coordinated interplay of scene, utterance, and world knowledge: evidence from eye tracking. *Cognitive Science.*

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.

Magliano, J. P., & Schleich, M. C. (2000). Verb aspect and situation models. *Discourse Processes*, *29*, 83–112.

Richardson, D. C., & Spivey, M. J. (2000). Representation, space and hollywood squares: Looking at things that arent there anymore. *Cognition*, *76*, 269–295.

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

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

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

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.

Yu, C., Ballard, D. H., & Aslin, R. N. (2005). The role of embodied intention in early lexical acquisition. *Cognitive Science*, *29*, 961–1005.

Zwaan, R. A., Maaden, C., & Whitten, S. N. (2000). The presence of an event in the narrated situation affects its availability to the comprehender. *Memory and Cognition*, *28*, 1022–1028.