Incremental Effects of Mismatch during Picture-Sentence Integration: Evidence from Eye-tracking

Pia Knoeferle (knoeferle@coli.uni-sb.de)
Department of Computational Linguistics,
Saarland University, 66041 Saarbrücken, Germany

Matthew W. Crocker (crocker@coli.uni-sb.de)
Department of Computational Linguistics
Saarland University, 66041 Saarbrücken, Germany

Abstract

A model of sentence-picture integration developed by Carpenter and Just (1975) predicts that picture-sentence integration ease/difficulty depends on picture-sentence match/mismatch respectively. Recent findings by Underwood, Jebbet, and Roberts (2004), however, fail to find a match/mismatch difference for serial picture-sentence presentation in a sentence verification study. In a sentence comprehension study with serial picture-sentence presentation we find no match/mismatch effect in total sentence inspection times. However, inspection times for individual sentence regions reveal a mismatch effect at the very sentence constituent for which the corresponding picture constituent mismatches, and this in a study with a sentence comprehension rather than verification task. Drawing on insights about spoken sentence comprehension during the inspection of concurrent scenes, we suggest that the absence of a mismatch effect in the Underwood et al. studies might be due to grain size of gaze time analyses.

Introduction

How do we integrate what we see in a scene with a sentence that we read? Answering this question is of interest in various types of comprehension situations such as when we read comic books (Carroll, Young, & Guertin, 1992), newspaper advertisements (Rayner, Rotello, Stewart, Keir, & Duffy, 2001), or inspect scientific diagrams (Feeney, Hola, Liversedge, Findlay, & Metcalf, 2003).

One account of how a picture and sentence are integrated is the "Constituent Comparison Model" (CCM) by Carpenter and Just (1975). They suggest that people build a mental representation of sentence and picture constituents, and that the corresponding constituents of sentence and picture are then serially compared with one another. Their model of sentence verification accounts for response latencies in a number of sentence-picture verification studies by attributing differences in the response latencies to congruence/incongruence between sentence and picture (e.g., Gough, 1965, Just & Carpenter, 1971).

In a sentence verification task, Just and Carpenter (1971) presented people with a picture of either red or black dots, followed by a related written sentence. Sentence verification response latencies were shorter when the colour adjective in the sentence (*red*) matched the colour of the depicted dots (red) than when it did not match their colour (black). The CCM predicts precisely that when there is a match between a picture and a sentence, their integration should be faster than when a picture and a sentence do not match.

A Model of Incremental Sentence-Picture Comparison?

The CCM has received strong support from off-line response latencies in verification tasks, and is primarily a model of sentence-picture verification. The model specifies - at least to some extent - how the integration of picture and a written sentence proceeds *incrementally*: by serially comparing the representations of sentence and corresponding picture constituents.

Reaction times are approriate for testing the complexity of sentence-picture integration steps. However, for truly examining the incremental integration of picture- and sentence-based mental representations they are less informative than other, on-line measures such as eye-tracking. In sentence-picture integration research, few studies have monitored eye-movements during sentence reading (e.g., Carroll et al., 1992; Underwood et al., 2004). Among recent studies in the sentence-picture verification paradigm that have employed eye-tracking, findings by Underwood et al. (2004) have challenged the validity of the CCM. They have further identified important additional factors (e.g., order of picture-sentence presentation) that affect their integration.

In two eye-tracking studies with a sentence-picture verification task, Underwood et al. (2004) examined the effect of presentation order for real-world photographs and captions. They report total inspection time, number of fixations, and durations of fixations for the entire sentence and picture in addition to response latencies. In Experiment 1, picture and caption were presented together, and congruence was manipulated (match/mismatch). Results confirmed the established match/mismatch effect: Response latencies were longer for the mismatch than for the match condition. Total inspection times and number of fixations further confirmed this finding.

In Experiment 2, order of presentation (picture-first, sentence-first) was introduced as a condition in addition to the match/mismatch manipulation. Crucially, and in contrast to Experiment 1, there was no match/mismatch effect in Experiment 2 in either response latencies or inspection times for the entire sentence. Response accuracy was relatively high (83.6 and 79.2 percent for match

and mismatch responses respectively, with no reliable difference between match/mismatch conditions). Further findings were faster reaction times for the sentence-first in comparison with the sentence-last condition. The reaction time findings were confirmed by total inspection times.

The findings by Underwood et al. challenge the general validity of the CCM as a model of sentence-picture integration. They lend support to the view that the sentence-picture integration process cannot be explained by one factor (e.g., match/mismatch), but might rather be the product of complex interactions between a number of factors.

Despite their merits in advancing our knowledge of sentence-picture integration, existing eye-tracking studies in the sentence verification paradigm (e.g., Carroll et al., 1992; Underwood et al., 2004) offer limited insights into how picture and sentence are integrated on a constituent-by-constituent basis. Underwood et al., report, for instance, only total reading times for the entire sentence rather than reading times for individual sentence regions.

It is gaze times in individual sentence regions, however, which prior psycholinguistic research on sentence processing has established as a reliable measure for the incremental processing of relevant syntactic, semantic, and pragmatic information (see Rayner, 1998 for review). We suggest that a more fine-grained analysis of gaze-durations may reveal more about the incremental nature with which sentence and picture are integrated when they match and when they are incongruent.

Does the Model Generalize Across Tasks?

An important issue for on-line sentence comprehension concerns the claims that the CCM makes in situations where the sentence-picture integration process is not modulated by a verification task. Serial picture-sentence constituent comparison is predicted to take place "whenever a person answers a question, follows an instruction, or incorporates statements into his belief system" (Carpenter & Just, p. 72).

Recall that there was no match/mismatch effect in the serial-order presentation study by Underwood et al. (2004). Underwood and colleagues suggest that its absence may be due to task requirements. They reason that in other studies (e.g., Goolkasian, 1996), the match/mismatch effect was only present when participants were asked to make a verbatim comparison between picture and words. If task requirements were indeed responsible for the presence or absence of the match/mismatch effect, then this would seem to question the general validity of the CCM as a model of sentence-picture processing.

To further investigate issues of task-specificity, let us consider sentence-picture integration in passive sentence comprehension and act-out tasks. While few psycholinguistic studies on on-line sentence comprehension have examined the integration of a *written* sentence and a picture (e.g., Carroll et al., 1992), a number of studies have monitored attention in scenes to investigate the interaction of scene and *spoken* sentence.

Studies using concurrent scene-utterance presentation, have revealed important insights into the incremental integration of an immediate scene context with an unfolding spoken sentence. Shortly after a word in the utterance identifies a scene constituent as its referent, people inspect the relevant scene object (e.g., Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). Sedivy, Tanenhaus, Chambers, & Carlson (1999) show the incremental influence of contrast between objects in the scene on semantic interpretation. Knoeferle et al. (2005) have in turn found a rapid influence of the immediately depicted scene events on on-line thematic role assignment once the verb in the sentence had identified the relevant depicted action and its associated role relations.

The close time-lock of utterance and attention in a scene has importantly been extended to serial picture-utterance presentation. In Altmann (2004) people inspected an image with a man, a woman, a cake, and a newspaper, and then the screen went blank (see also Richardson & Spivey, 2000; Spivey, Richardson, & Fitneva, 2004). Two seconds later, people heard a sentence that described part of the previously-inspected scene (e.g., *The man will eat the cake*). Once people had heard the verb in the sentence, they rapidly looked at the location on the blank screen where previously there had been a cake. The time-course of eye-movements in the serial-presentation study closely ressembled the time-course of gaze-patterns in an earlier study (Altmann & Kamide, 1999) with concurrent scene and utterance presentation. These findings are consistent with the CCM. The close time-lock between utterance comprehension and scene inspection observed in all of these studies reflects a serial comparison and integration of the unfolding utterance constituents and corresponding scene constituents. The findings from spoken sentence comprehension in scenes even go beyond the serial constituent-by-constituent integration predicted in the CCM and show a certain degree of anticipation in sentence-picture integration. The important insight from the above studies for the present paper is, however, that a serial picture-utterance integration was observed during both concurrent and serial picture-utterance presentation.

The present paper directly explores whether the CCM is valid for serial presentation, whether it generalizes from sentence verification to a sentence comprehension task, and whether a more fine-grained analysis of the eye-gaze data provides more detailed insights into the incremental integration of picture and written sentence for serial picture-sentence presentation. To this end, we designed a study which combined a sentence comprehension task (e.g., Altmann & Kamide, 1999; Knoeferle et al., 2005) and a match/mismatch manipulation during serial presentation.

If serial presentation and/or task eliminates the match/mismatch effect, then we should see no such effect in the eye-gaze data of our study. Finding a difference, however, in sentence or word-region gaze-times would support the view that the Carpenter and Just

model does generalize to tasks other than verification and across presentation order. Furthermore, if the grain-size of gaze-analyses is one reason why Underwood et al. (2004) failed to find a difference between picture-sentence integration in the match/mismatch conditions, then we would expect to observe a difference in finer-grained word-region gaze-times while replicating their finding that total sentence inspection times showed no effect of match/mismatch.

To investigate in addition the influence of other factors on sentence-picture integration, we manipulated the word order of the sentence. In this respect, the present picture-written sentence study continues research by Knoeferle et al. (2005) who investigated comprehension of spoken German sentences with local structural and thematic role ambiguity. German is a language with relatively flexible word order. Both a subject-verb-object (SVO), and an object-verb-subject (OVS) ordering of constituents is grammatical, with the SVO order being preferred. People inspected a scene in which a princess both paints a fencer and is washed by a pirate while hearing either an initially ambiguous SVO or OVS sentence. Late disambiguation occurred through case-marking on the second noun phrase, and earlier disambiguation was only possible through depicted events. When the verb identified an action, its associated depicted role relations disambiguated towards either an agent-patient (SVO) or patient-agent role relation (OVS), as evidenced by anticipatory eye-movements to the patient (pirate) or agent (fencer) respectively. A further goal of the present study is thus to investigate the integration of depicted event scenes and reading mechanisms for initially ambiguous written German sentences.

Experiment

Method

Participants Thirty-six German native speakers with normal or corrected-to-normal vision received each 7.50 euro for taking part in the experiment.

Materials We created 108 images using commercially available clipart and graphics programs. An image either depicted a female-action-male (e.g., granny-filming-businessman, Fig. 1a), a male-action-female event (businessman-filming-granny, Fig. 1b), or there was no explicitly depicted (granny, businessman, Fig. 1c). There were 72 sentences, 36 of which described a female-action-male, and 36 of which described a male-action-female event. The experiment was carried out in German. Since the feminine character was always mentioned first, and the first noun phrase was case- and role-ambiguous, sentences were locally structurally ambiguous. One sentence of an item had a subject-verb-object (SVO) order (see Table 1), and the second sentence always had an object-verb-subject (OVS) order. The only difference between the two sentences was the disambiguating definite determiner of the second noun phrase, which was in the accusative case (*den*, 'the') for SVO, and in the nominative case (*der*, 'the') for OVS sentences.

Design From the 108 images and 72 sentences, we created 36 item sets. An item consisted of three images (e.g., Fig. 1) and two written sentences (e.g., Table 1). A three-by-two within-subject design crossed *congruence* (match, mis1, mis2) with *sentence type* (SVO, OVS), resulting in six conditions. Each of the two sentences in Table 1 was presented with each of the images (a, b, and c in Fig. 1).

Table 1: Example Item Sentences

Cond.	Sentence
SVO	Die Oma (subj.) filmt soeben den Handelskaufmann (obj.) nach dem Vertragsabschluss. The granny (subj.) films currently the businessman (obj.) after the signing of the contract. 'The granny films currently the businessman after the signing of the contract.'
OVS	Die Oma (obj.) filmt soeben der Handelskaufmann (subj.) nach dem Vertragsabschluss. The granny (obj.) films currently the businessman (subj.) after the signing of the contract. 'The businessman films currently the granny after the signing of the contract.'

When the SVO sentence was presented with Fig. 1a, it matched the depicted agent-patient role relations (granny-filming-businessman) in the scene (SVO-match). In contrast when it was presented with Fig. 1b, the scene contained constituents that matched individual sentence constituents (NP1, verb, NP2), however, the role relations in the scene (businessman-filming-granny) were the opposite of the thematic relations expressed by the SVO sentence in Table 1 (SVO-mis1). When presented with Fig. 1c, the scene did not contain explicitly depicted actions, and was incongruent with the sentence through an absence of the thematic relations expressed by the sentence (SVO-mis2). When the OVS sentence in Fig. 1, appeared together with Fig. 1a the thematic relations in the sentence (businessman-filming-granny) were the opposite of the role relations in the scene (granny-filming-businessman) (OVS-mis1). In contrast, presenting the OVS sentence with Fig. 1b resulted in a match of depicted and thematic role relations (OVS-match). For Fig. 1c, the scene was incongruent with the sentence since it did not explicitly depict role relations (OVS-mis2).

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. Pictures and sentences were presented serially, and the picture was always displayed first. For each trial, participants first focussed a centrally-located fixation dot on the screen. The experimenter then triggered the image, which participants inspected. Once they had inspected and understood the picture, they indicated this by pressing a button. A template with a

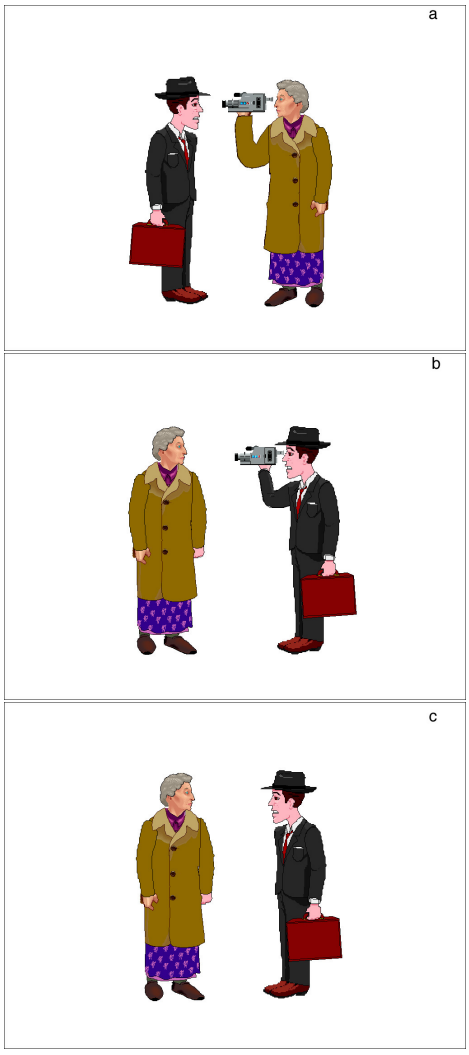

Figure 1: Example Item Images

black fixation square at the position of the first word in the subsequent sentence appeared automatically. Once participants fixated the square, the experimenter presented the sentence. Participants read the sentence, and indicated that they had read and understood it by pressing a button. The only task was to read and understand the sentences, i.e., there was no explicit sentence verification task. There were no questions on any of the images, and no questions on the item sentences. However, for 24 of the 48 filler trials participants answered a yes/no question which always referred to the sentence.

The items were randomized and one version of each item was assigned to one of six experimental lists. Items were rotated across lists such that an equal number of each condition (SVO-match, SVO-mis1, SVO-mis2, OVS-match, OVS-mis1, OVS-mis2) appeared in each list and such that no participant saw more than one version of each item. Each list consisted of 36 experiment and 48 filler items. Consecutive experiment trials were separated by at least one filler item. The entire experiment lasted approximately 45 min with a short break after half the trials.

Results

Figs 2, 3, 4, and 5 show the mean total reading times (duration of all fixations in a region) per condition for the NP1 (e.g., 'The granny'), verb ('films'), adverb ('currently'), and NP2 ('the businessman') sentence regions respectively. For the inferential analyses, we carried out repeated measures ANOVAs. Analysis by participants are reported as F1, and by items as F2.

There was no effect of congruence (match, mis1, mis2) on the NP1 region in total time (Fig. 2), both $Fs < 1$, nor was there an interaction between congruence and sentence type. We found an effect of sentence type on NP1, $F1(1, 35) = 12.39, p < 0.01, F2(1, 35) = 18.89, p < 0.001$.

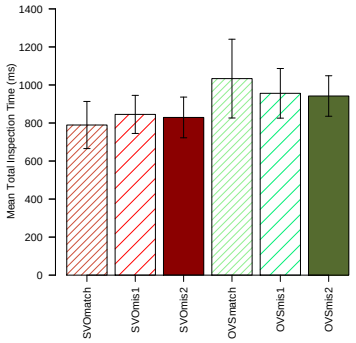

Figure 2: Mean total reading time for the NP1 region

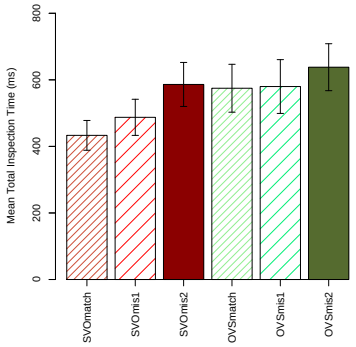

Figure 3: Mean total reading time for the verb region

For the verb region, where the mismatch occurred (Fig. 3), in contrast, the key finding was a main effect of congruence (match, mis1, mis2), $F1(2, 34) = 16.51, p < 0.001, F2(2, 34) = 7.22, p < 0.01$. There was further a main effect of sentence type (SVO, OVS) on the verb,

$F1(1,35) = 17.19, p < 0.001, F2(1,35) = 23.40, p < 0.001$, in the absence of an interaction between congruence and sentence type, $ps > 0.2$.

In the adverb region (Fig. 4), total reading time analyses revealed a main effect of congruence by participants (match, mis1, mis2), $F1(2,34) = 3.37, p < 0.05, F2(2,34) = 2.69, p = 0.08$. There was further a main effect of sentence type (SVO, OVS), $F1(1,35) = 18.29, p < 0.001, F2(1,35) = 40.28, p < 0.001$, and a marginal interaction of congruence and sentence type, $F1(2,34) = 2.75, p = 0.08, F2(2,34) = 1.52, p = 0.23$.

For the NP2 region (see Fig. 5) analyses revealed only a main effect of sentence type, $F1(1,35) = 44.12, p < 0.001, F2(1,35) = 90.54, p < 0.001$, no effect of congruence, and no interaction between sentence type and congruence (all $ps > 0.1$).

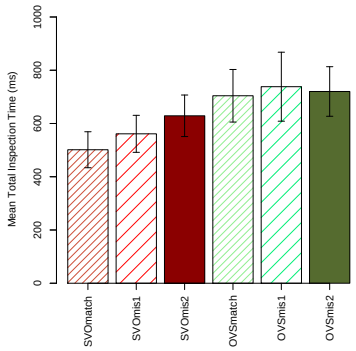

Figure 4: Mean total reading time for the adverb region

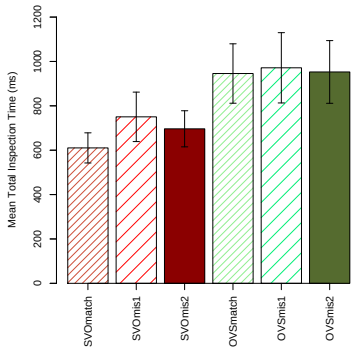

Figure 5: Mean total reading time for the NP2 region

The main effect of congruence (match, mis1, mis2) that appeared in total reading times on the verb region, was also apparent in the first fixation measure on the adverb region, $F1(2,34) = 4.99, p < 0.02, F2(2,34) = 6.64, p < 0.01$ ($Fs < 1$ for congruence effects on the other analysis regions for first fixation).

In contrast to the analyses for the individual regions, neither the duration of inspection for the whole sentence $Fs < 1$, nor mean total sentence inspection time ($ps > 0.2$) showed a reliable match/mismatch effect.

Discussion

The eye-gaze data that we report provide strong support for the Constituent Comparison Model, and furthermore allow us to integrate predictions by the model with prior research on picture-sentence integration during concurrent scene and utterance presentation (Knoeferle, 2004, Knoeferle & Crocker, 2004; Knoeferle et al., 2005; Sedivy et al., 1999; Tanenhaus et al., 1995).

The fact that we observed a match/mismatch effect in a sentence comprehension task shows that the CCM generalizes to tasks other than verification, and supports its validity as a model of picture-sentence integration. In more detail, analyses of gaze durations (at the verb) revealed a match/mismatch effect and that even for a presentation type (serial picture-sentence presentation) for which Underwood et al. (2004) had failed to find a match/mismatch effect in their gaze measures. Reading times on the verb were highest when the scene did not explicitly depict the thematic relations described in the sentence (mis2), lowest when the role relations depicted in the scene matched those described in the sentence (match), and intermediate, when scene relations were opposite to the thematic relations expressed in the sentence (mis1). The picture-sentence match/mismatch effects that we observed on the very constituent (the verb) at which the mismatch occurred clearly support the CCM prediction that picture-sentence integration is a process of serially comparing sentence constituents with mental representations of the corresponding picture parts.

Crucially, while fine-grained, constituent-based readings times in our experiment did reveal a significant effect of match/mismatch, we - just as Underwood et al. (2004) - failed to find an effect of match/mismatch in total sentence inspection times. Our findings therefore suggest that the lack of a mismatch as reported by Underwood et al. (2004) derives from their reliance on total sentence times rather than on finer, constituent-based analyses of the sentence.

Perspectives on Scene-Sentence Integration

Findings from the present study lend strong support to the CCM, while also providing an explanation for why Underwood et al. failed to find confirming support. The CCM specifies two key steps in the integration process. After the acquisition of mental representations from scene and sentence, the second step concerns the constituent-based comparison and integration of the acquired representations. Indeed, it was precisely the constituent-by-constituents analyses of total gaze durations which revealed the mismatch effect, total sentence times, as reported by Underwood et al., did not.

The success of the CCM in explaining our findings despite the lack of an explicit verification task in our study leads us to speculate about the applicability of CCM as a more general model of scene-sentence comprehension.

Closer consideration of the CCM suggests that the mechanisms which it proposes have been heavily shaped by the nature of the verification paradigm within which the model has been developed. Crucially, we suggest that full scene-sentence comprehension goes beyond the two CCM steps, and crucially relies upon interactive and interpretative processes. This is particularly evidenced by the studies on spoken comprehension of Tanenhaus et al. (1995), Sedivy et al. (1999), and Knoeferle et al. (2005). These studies reveal the dynamic influence of the scene on utterance comprehension processes which manifests itself by anticipatory eye-movements to likely scene elements, while or even before they are mentioned.

An account of scene-sentence integration which is compatible with the CCM, but which furthermore includes such interpretative processing has been proposed by Knoeferle & Crocker (2004) based on findings from utterance comprehension during scene inspection. Their "Coordinated Interplay Account" describes scene-sentence interaction as highly incremental, with a close time-lock between scene and sentence processing. Importantly, it further predicts that once the utterance has identified relevant scene information (e.g., a depicted action), that scene information can then in turn influence comprehension and interpretation processes - a process which is missing from the Constituent Comparison Model. This last step, from a simple incremental integration of sentence and scene towards an interactive and interpretative account of these comprehension processes is an important first step towards a complete and general model of comprehension in visual environments.

Acknowledgements

We thank Martin J. Pickering for his helpful comments, and Nicole Kühn for her assistance in running and analysing the experiment. This research was funded by a PhD scholarship to the first, and by SFB 378 project ALPHA to the second author, both awarded by the German Research Foundation (DFG).

References

Altmann, G. T. M. (2004). Language-mediated eye movements in the absence of a visual world: the blank screen paradigm. *Cognition*, *93*, B79-B87.

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

Carpenter, P. A. & Just, M. A. (1975). Sentence comprehension: a psycholinguistic processing model of verification. *Psychological Review*, *82*, 45-73.

Carrol, P. J., Young, J. R., & Guertin, M.S. (1992). Visual analysis of cartoons: A view from the far side. In K. Rayner (Ed.), *Eye movements and visual cognition: Scene perception and reading* (pp. 444-461). NewYork: Springer-Verlag.

Feeney, A., Hola, A. K. W., Liversedge, S. P., Findlay, J. M., & Metcalf, R. (2000). How people extract information from graphs: Evidence from a sentence-graph verification paradigm. In M. Anderson, P. Cheng, & V. Haarslev (Eds.), *Theory and application of diagrams: First international conference, Diagrams 2000* (pp.149-161). Berlin: Springer-Verlag.

Goolkasian, P. (1996). Picture-word differences in a sentence verification task. *Memory & Cognition*, *24*, 584-594.

Gough, P. B. (1965). Grammatical transformations and speed of understanding. *Journal of Verbal Learning and Verbal Behavior*, *5*, 107-111.

Just, M. A., & Carpenter, P. A. (1971). Comprehension of negation with qualification. *Journal of Verbal Learning and Verbal Behavior*, *10*, 244-253.

Knoeferle (2004). The role of visual scenes in spoken comprehension. Unpublished doctoral dissertation, Saarland University.

Knoeferle & Crocker (2004). The coordinated processing of scene and utterance: evidence from eye-tracking in depicted events. In: *Proceedings of the International Conference on Cognitive Science*, (pp. 217-222), Allahabad, India.

Knoeferle, Matthew Crocker, Martin Pickering, & Christoph Scheepers (2005). The influence of the immediate visual context on incremental thematic role-assignment: evidence from eye-movements in depicted events. *Cognition*, *95*, 95-127.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*, 372-422.

Rayner, K., Rotello, C. M., Stewart, A. J., Keir, J., & Duffy, S. A. (2001). Integrating text and pictorial information: Eye movements when looking at print advertisements. *Journal of Experimental Psychology: Applied*, *7*, 219-226.

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

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

Spivey, M. J., Richardson, D. C., & Fitneva, S. A. (2004). Thinking outside the brain: Spatial indices to visual and linguistic information. In J. M. Henderson, & F. Ferreira (Eds.), *The interface of language, vision, and action: Eye movements and the visual world*. New York: Psychology Press.

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.

Underwood, G., Jebbett, L., & Roberts, K. (2004). Inspecting pictures for information to verify a sentence: eye movements in general encoding and in focused search. *The Quarterly Journal of Experimental Psychology*, *56*, 165-182.