Incremental Syntactic Disambiguation Using Depicted Events: Plausibility, Co-Presence and Dynamic Presentation

Emilia Ellsiepen (ellsiepe@coli.uni-sb.de)
Department of Computational Linguistics,
Saarland University, 66041 Saarbrücken, Germany

Pia Knoeferle (pknoeferle@uscd.edu)
Center for Research in Language,
University of California, San Diego, USA

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

Abstract

Prior research in the visual world paradigm has shown that co-present depicted events rapidly influence on-line structural disambiguation. The generality of this finding, however, is uncertain, since these studies relied on the visual co-presence of the depicted events, and on highly non-stereotypical stimuli which may have increased the salience of visual information. We conducted three visual world experiments to address these concerns. The first experiment used more plausible stimuli than prior studies. The second experiment investigated the disambiguation effect for prior events using the "blank screen" paradigm. In the third experiment, actions were animated and completed before the utterance started. While a disambiguation effect was found in all three experiments, the first and second showed an additional speed-up compared to previous studies while the third suggests competition with visual factors.

Introduction

Existing research shows that language is not processed in isolation, but that scene information is used very rapidly to inform online sentence comprehension. Spoken language has also been shown to guide visual attention in a scene. In particular, eye-tracking experiments conducted in the visual world paradigm provided evidence for the use of a visual referential context for lexical access (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995; Allopenna, Magnuson, & Tanenhaus, 1998), anticipation of thematic role fillers present in the visual scene (Altmann & Kamide, 1999; Kamide, Scheepers, & Altmann, 2003), and structural disambiguation using a referential context (Tanenhaus et al., 1995) and depicted events (Knoeferle, Crocker, Scheepers, & Pickering, 2005). The rapidity with which scene information is used and the fact that it influences not only semantic interpretation but crucially core linguistic processes such as structural disambiguation point to the fundamental importance of scene information for incremental situated comprehension. This claim receives further support from eye-tracking results suggesting that scene information may be prioritized for comprehension over stereotypical thematic role knowledge (Knoeferle & Crocker, 2006).

For the findings of Knoeferle et al. (2005) and Knoeferle and Crocker (2006), it is not clear to what degree the importance of scene information was accurately reflected by these studies, since they relied on the visual co-presence of the depicted events, and on highly non-stereotypical stimuli which may have increased the salience of and reliance upon visual information. We conducted three visual world experiments to address these concerns with regard to the use of depicted events for syntactic disambiguation.

Limitations of previous studies

While Knoeferle et al. (2005) were able to show that listeners exploited information provided by depicted events in a visual scene for syntactic disambiguation of locally ambiguous subject-verb-object (SVO) and object-verb-subject (OVS) German sentences, several aspects of their experiments limit the generality with which we might interpret these findings. In particular, implausibility of scenes and sentences, co-presence of scene and utterance, and static presentation of events might have biased listeners to rely more on scene information than they naturally would.

The scenes in their study differed from real scenes in everyday life in at least three respects. Firstly, characters performed actions they were unlikely to perform in a natural context to avoid stereotypical relations between verbs and nouns, e.g. a princess washing somebody. Secondly, the character constellations were highly implausible in that there is no natural context in which these would occur together, e.g. a princess, a pirate, and a fencer. Thirdly, some characters were historical or fictional like an amazon or a mermaid, so that participants would not have seen them or interacted with them in real life. Due to these implausibilities, listeners might have been biased to rely more on the visual information, including the depicted events. If so, scene information may have had a disproportionately strong influence, especially since participants could not rely on their own experience with events to the same extent as they might have done in real-world environments where events are typically plausible.

In a natural context, utterance and corresponding scene are often not available at the same time. While it has been shown that a previously viewed scene can lead to anticipation of role fillers in unambiguous sentences (Altmann, 2004; Knoeferle & Crocker, 2007), the influence of a prior scene on disam-

Table 1: Example item

Image	Condition	Sentence
1 A	SVO	Das Kindergartenkind malt gerade den Onkel. The small child (ambiguous) paints currently the uncle (object). 'The small child is currently painting the uncle.'
1 A	OVS	Das Kindergartenkind haut gerade der Knabe. The small child (ambiguous) hits currently the boy (subject). 'The boy is currently hitting the small child.'
1 B	SVO	Das Kindergartenkind haut gerade den Knaben. The small child (ambiguous) hits currently the boy (object). 'The small child is currently hitting the boy.'
1 B	OVS	Das Kindergartenkind malt gerade der Onkel. The small child (ambiguous) paints currently the uncle (subject). 'The uncle is currently painting the small child.'

biguation processes has not yet been established. If it can be shown also for this core syntactic process, it would considerably strengthen the claim that the use of scene information in language processing is not dependant on co-presence.

Furthermore, action events are often not static, but unfold over time. It is therefore interesting to see, whether disambiguation takes place when events are presented dynamically and prior to utterance presentation such that a dynamic and completed, rather than a static, event has to be kept in working memory. Knoeferle and Crocker (2007) found that animated actions that end before the utterance starts can still be used in role filler anticipation, but that they do not take priority over stereotypical role filler knowledge any longer. This suggests that event information is less accessible after the event ended and might not be exploited as easily for syntactic disambiguation as information from static events.

We present three experiments that address these issues. For Experiment 1, scenes were constructed that are plausible with regard to the above mentioned aspects. Experiment 2 makes use of the "Blank Screen Paradigm" (Richardson & Spivey, 2000; Altmann, 2004), where the picture is replaced by a blank screen before the utterance starts. In Experiment 3, events are presented dynamically, as a sequence of 9 frames, before sentence onset.

Experiment 1

This experiment investigates the use of plausible depicted events in syntactic disambiguation of locally ambiguous sentences. Knoeferle et al. (2005) found a disambiguation effect as evidenced by anticipatory looks to the depicted character corresponding to the still missing role filler shortly after the verb had identified a depicted event in the scene as relevant, i.e. on the post-verbal adverb. Thus, if people use the comparatively more plausible events in our study for disambiguation in real time, then we should see a similar disambiguation effect in the gaze pattern for our study. If, however, Knoeferle et al. (2005) only found the effect because the events were

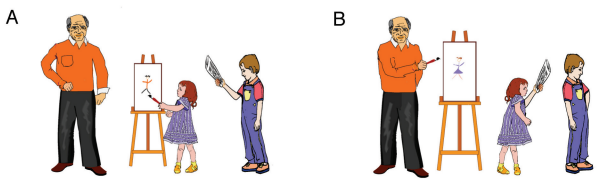

Figure 1: Example picture with counterbalancing version

so non-stereotypical, attracting a disproportionate amount of visual attention, disambiguation with more plausible and possibly less salient events may be delayed. In this case, we may find no disambiguation effects prior to the second noun phrase (NP2) that disambiguates through unambiguous case marking. A third possibility is that the added effects of plausibility facilitates scene-sentence integration, and thus accelerate the disambiguation effect. If this were the case, then we might expect to see an even earlier disambiguation effect than Knoeferle et al. (2005), i.e., beginning during the verb rather than for the post-verbal time window.

Method

Participants Thirty-two students from Saarland University were paid 5 Euro each to take part in the experiment. All of them were native speakers of German and had normal or corrected-to-normal vision.

Materials Sixteen items were created. Each consisted of two pictures (one counterbalancing) and four recorded German sentences, structurally ambiguous up to and including the postverbal adverb. A complete example can be found in Figure 1 and Table 1.

Each item picture showed three characters involved in two agent-action-patient events with one character being ambiguous, i.e. agent of one action and patient of the other. The other two were either agent or patient of one of the actions. They will be referred to as 'agent', 'ambiguous character' and 'pa-

tient' from now on. The two pictures of the same item were different only in the direction of the actions to counterbalance possible plausibility biases. While strongly stereotypical relations (e.g. a wizard jinxing somebody) were avoided, the character combinations in a scene (e.g. small child, boy, uncle) and actions (e.g.'small child paints uncle', 'boy hits small child') were chosen to be plausible.

Each item was shown in two conditions (see table 1). In both conditions, the initial NP was ambiguously case-marked as either subject or object and referred to the ambiguous character. In one condition, the sentence was in SVO with the ambiguous character as subject, in the other condition, it was in OVS with the ambiguous character as object. Nouns referring to agent and patient were matched in length within an item, and for frequency of lemmas between all items using CELEX (Baayen, Piepenbrock, & Gulikers, 1995). Since prosody can affect early disambiguation between OVS and SVO (Weber, Grice, & Crocker, 2006), for half of the items the OVS and SVO recordings were cross-spliced up to and including the adverb to avoid any possible influence of intonation. Additionally, the speaker was instructed to keep the intonation neutral.

In addition to the 16 experimental items, there were 32 filler trials. They were designed to introduce more linguistic and visual variability, to distract participants from the purpose of the experiment, and prevent them from developing strategies. The fillers contained generally plausible sentences and scenes like the experimental items.

Four lists were created, each containing every item in only one condition and 32 fillers. Lists were pseudo-randomized individually for every subject with three fillers in the beginning, and at least one filler trial separating experimental trials.

Procedure Participants' eye-movements were monitored with a sampling rate of 250 Hz using an SMI EyeLink head-mounted eye-tracker. Pictures were presented on a 21" color monitor at a resolution of 1024x768 pixels, sentences were played over loudspeakers. The picture was presented for a total of 7000 ms with the sentence starting after 1000ms of preview time. At the end of the 7000 ms the trial was terminated automatically. There was no other task than to observe the presented pictures, to pay attention to the spoken sentences, and to try to understand both sentence and scene as well as possible. The experiment lasted approximately 30 min.

Analysis Fixations were related to object regions in the pictures. This was done with the use of image templates with the four regions 'agent', 'patient', 'ambiguous character' and 'background', the regions exceeded the original space occupied by the characters by about 30pixels, except the background. Subsequent fixations on the same region were pooled into inspections. We report the time course of inspection proportions in the form of time curves. In addition, we conducted repeated measure ANOVAs on the summed inspection durations. For this, inspections were related to different time windows, which are relative to the exact word boundaries of each

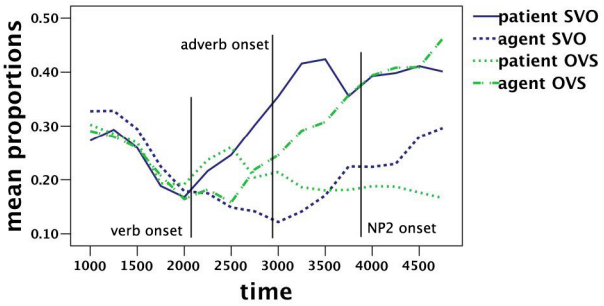

Figure 2: Time course of look proportions Experiment 1

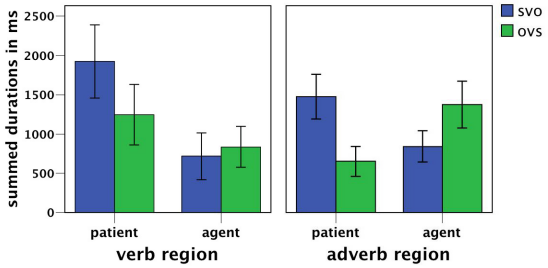

Figure 3: Summed inspection durations Experiment 1

individual trial. The regions used for analysis were Verb and Adverb, both stretching from word onset to the onset of the next word. Only inspections which start in the time window were counted. For the Adverb region, the parts of inspections falling after the onset of the second NP, which disambiguated between the two structures were excluded. The two factors included in the ANOVAs are *structure* (OVS, SVO) and *character* (agent, patient, ambiguous and background). Besides an omnibus ANOVA including all levels of both factors, a follow-up ANOVA only including agent and patient in *character* was conducted. In addition to the ANOVAs, planned pairwise comparisons were conducted between the factor combinations agent-OVS and patient-OVS, agent-SVO and patient-SVO, agent-OVS and agent-SVO, and patient-OVS and patient-SVO. The significance level was adjusted using the Bonferroni correction.

Results and Discussion

The time course of inspection proportions (Fig.2) reveals a rise in looks towards the patient in both conditions right after verb onset. For the SVO condition inspections of the patient continue for the rest of the utterance. In OVS sentences, however, looks to the agent increase, exceeding looks to the patient before the onset of the adverb, continuing throughout the utterance. This indicates a disambiguation effect already during the verb. The cause of the early fixation of the patient in both conditions likely reflects the fact that the first mentioned character is oriented towards the patient and performs an action on him. In a canonical SVO sentence, the patient would be mentioned next.

The results from the inferential analysis of summed look durations (Fig.3) confirm this time curve pattern: We find significant interactions between *character* and *structure* for both, Verb ($F1(2, 58)=6.4$ $p<.005$, $F2(2, 31)=3.4$ $p<.05$) and Adverb region ($F1(2,64)=13.1$ $p<.001$, $F2(2,30)=14$ $p<.001$). At the same time, we see a main effect of *character* for the Verb region ($F1(2, 55)=30.1$ $p<.001$, $F2(1, 24)=33.1$ $p<.001$), due to longer inspections of the patient.

Paired comparisons revealed that during the verb looks to the patient in SVO condition were significantly longer than looks to the agent in SVO condition as well as to the patient in OVS. Looks to the agent in OVS, however, were not significantly different in duration from looks to the patient and looks to the agent in SVO. In the Adverb region on the other hand, all paired comparisons yielded significant differences. This indicates a different time course of disambiguation for the two structures: In the canonical SVO structure anticipation of the correct role filler emerges slightly earlier than in the OVS condition.

In brief, we found evidence for structural disambiguation through depicted events in both the time curve pattern and analyses of inspection durations. The overall gaze pattern replicates the findings by Knoeferle et al. (2005). Crucially, however, we find evidence for disambiguation one region earlier (at the verb) than they did. This is consistent with the prediction that plausible scenes may actually facilitate the use of scene events, a point we return to in the general discussion.

Experiment 2

This experiment examines structural disambiguation using recently inspected scenes. As noted above, it has been previously shown that co-presence is not necessarily a prerequisite for the use of scene information in language processing. In experiments in which participants inspected a scene that was removed prior to presentation of a related utterance, they still directed eye-movements to the regions where referents used to be, indicating that those referents were still accessible for language processing (Altmann, 2004; Knoeferle & Crocker, 2007). If this generalized to the use of previously viewed events for syntactic disambiguation, we would expect to see the same pattern as in Experiment 1, even when scenes are not co-present.

Method

Participants Thirty-two further students from the same population were paid 5 Euro each to take part in the experiment.

Materials, procedure, and analysis The materials from the first experiment were used. The procedure also remained the same except for the presentation of the trials. In this experiment first the picture was displayed for 6000 ms. Then the screen went blank and the sentence was played after a delay of 1000 ms. The trial ended automatically 6000 ms after the screen went blank, the total duration of one trial was thus 12000 ms. Subjects were told to observe and understand the presented scenes and pay attention to the spoken sentence. They were also made aware that eye movements were recorded both with the picture present and with the blank screen.

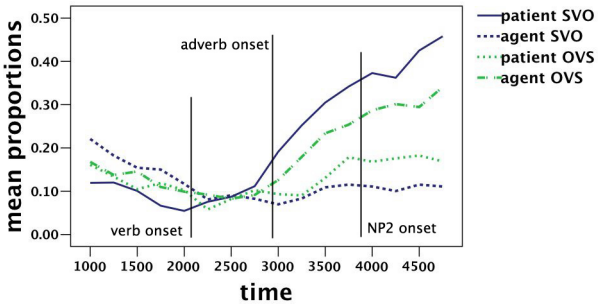

Figure 4: Time course of look proportions Experiment 2

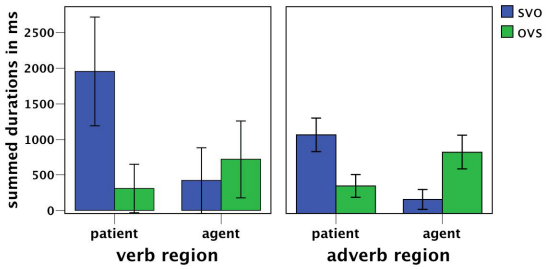

Figure 5: Summed inspection durations Experiment 2

The analysis was the same as for Experiment 1. There were, however, a number of trials excluded from the analysis because participants did not move their eyes when the utterance was spoken. All trials which had fixations only to ambiguous character and background were thus excluded from the analysis, 15% in total.

Results and Discussion

The overall pattern is similar to the results in Experiment 1 with some difference apparent in the time course. In Fig. 4, we see increased looks to the agent in OVS and to the patient in SVO before the onset of the adverb, continuing throughout the utterance, while looks to the irrelevant character, i.e. agent in SVO and patient in OVS, level off. This indicates ongoing disambiguation between the two structures during verb and adverb. The disambiguation effect is replicated also in the analysis of inspection durations (Fig.5): In SVO, participants looked longer on the patient during both, verb and adverb, while in OVS they looked longer on the agent during the adverb. These findings were confirmed by significant interactions in the ANOVAs for the Verb region($F1(3, 93)=3.73$ $p<.05$, $F2(2,35)=5.4$ $p<.01$)) and for the Adverb region ($F1(3, 75)=20.4$ $p<.001$, $F2(2, 29)=12.4$ $p<.001$).

Experiment 2 thus shows that events depicted in previously viewed scenes, still influence the core mechanism of syntactic disambiguation. We further replicate the early use of event

information in plausible scenes. In contrast to Experiment 1, however, the early fixations of the patient for both conditions right after verb onset are missing in the time curve. We already suggested this early effect may be due to the visual factor orientation, thus the absence of orientation information in the blank screen experiment would explain this difference.

Experiment 3

Experiment 3 investigates the influence of dynamic and completed events for syntactic disambiguation. With this manner of presentation, action events are perceived in a more similar way to a natural context, i.e. they are dynamic rather than static. Our aim is to establish firstly whether such dynamic presentation enables disambiguation, or whether the fact that events are completed may diminish their use. We also investigate whether recency affected the relative importance of completed events.

Method

Participants Thirty-two further students from the same population were paid 5 Euro each to take part in the experiment.

Materials and procedure To accomplish the impression of a fluid motion, seven pictures were composed for each item of Experiment 1. In one picture, all characters were in neutral positions with both events absent. Three pictures showed the unfolding of each action, with the other action absent. Pictures were then arranged in sequences: first the neutral version, than the three pictures of one action, again the neutral one, the three pictures of the second action, and, in the end, the neutral picture, resulting in 9 consecutive frames.

Recency was added as an additional factor, resulting in 4 conditions. In the recent condition, the event relevant for disambiguation was presented last, while in the not-recent condition, it was presented first.

The procedure remained the same, except the presentation of the trials. The picture sequences, lasting for 2600ms, were shown first. Then the neutral picture was displayed for a total of 6000ms with the sentence starting after 1000ms of silence. The total duration was thus 8600ms.

Results and Discussion

While a disambiguation effect was found, the results differ from those of the other two experiments in several respects. In the time course of look proportions (Fig.6), we see a bias towards fixating the patient in both conditions. Similar to Experiment 1, inspections of the patient increase right after verb onset, but in this experiment, they stay at a high level also in OVS. Inspections of the agent in OVS start to rise shortly before the onset of the adverb, but do not exceed the patient before the onset of the second NP. In the analysis of inspection durations (Fig.7), we do not find an interaction of *character* and *structure* in the Verb region and only a marginal effect for the Adverb region. In the post hoc defined combined disambiguation region stretching from verb onset to the onset

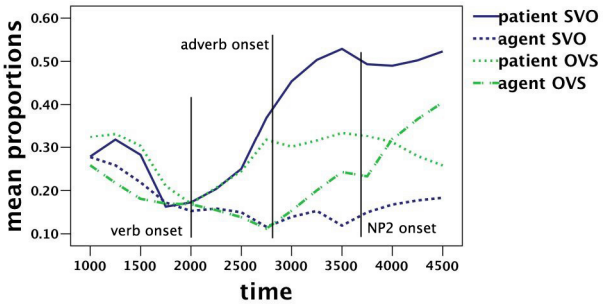

Figure 6: Time course of look proportions Experiment 3

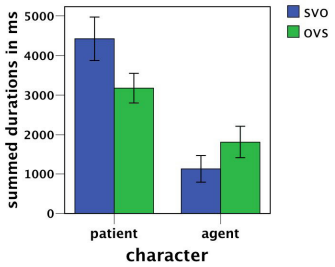

Figure 7: Inspection durations for the combined region of verb and adverb Experiment 3

of NP2, however, there is a significant interaction of *character* and *structure* ($F1(2, 63)=12.8$ $p<.001$, $F2(2, 26)= 13.6$ $p<.001$). Here, participants looked longer at the patient in SVO than in OVS and longer on the agent in OVS than in SVO. The factor *recency* had no effect in any of the analyses.

In sum, we continue to observe the influence of events for disambiguation, when they are dynamic, but this influence appears to be diminished. The strong bias towards fixating the patient could again be due to the orientation of the characters. Since the events were already completed when the utterance was played and the orientation was still present, event information apparently suffered a loss in salience.

General Discussion

The three experiments presented in this paper reveal the robust use of plausible scene events for syntactic disambiguation when scene and utterance were co-present, when the scene preceded the utterance, and with animated actions that were completed before the utterance started. This was evidenced by increasing anticipatory looks towards the agent character in OVS and the patient character in SVO as soon as listeners heard the verb: More inspections of the relevant character after verb onset were revealed in the time curves, and longer inspections in the inferential analysis during verb and adverb, before linguistic disambiguating information of case marking had been heard. Differences between the experiments were found in the time course of gaze behaviour. In Experiment 1, which used plausible scenes co-present with the utterance, the inferential analysis yielded significant inter-

actions of character and structure for both verb and adverb regions. For Experiment 2, where plausible scenes were shown prior to utterance presentation and then removed, anticipatory looks started to rise later, but before the onset of the adverb, with the inferential analysis still revealing an effect for both regions. In Experiment 3, plausible events were presented dynamically and were completed before the utterance started. Here the disambiguation effect was delayed: The patient was initially anticipated for both conditions. Looks to the agent for OVS increased only after adverb onset, and did not exceed looks to the patient until after the onset of the second NP, which disambiguates via case marking. In the inferential analysis, only the combined region of verb and adverb revealed a disambiguation effect, suggesting a diminished influence of event information.

Compared to the results of Knoeferle et al. (2005), we found an earlier disambiguation for plausible scenes: While they found the effect only at the adverb, Experiments 1 and 2 exhibit an interaction of *character* and *structure* already during the verb. This indicates that their results were not dependent on highly non-sterotypical scenes, and rather suggests that plausible event information can be used more effectively or at least more rapidly. An alternative explanation would be that the sentences themselves were easier to process because of their increased plausibiliy.

The results of Experiments 2 and 3 give insights into the details of the use of information from recently inspected events. The fact that disambiguation is delayed in both experiments supports the view that event information not copresent with the utterance cannot be accessed as easily as information immediately available in the scene, because it has to be retrieved from visual memory and possibly decays over time. More specifically, while in Experiment 1, there are early looks on the patient in both conditions, the same pattern is not found in Experiment 2, but much stronger again in Experiment 3. An explanation for those differences may be the competition between depicted event information and orientation of the ambiguous character towards the patient, and the difference in availability of the two information types: In Experiment 1, both sources are present at the same time and thus compete, leading to an early influence of orientation. In Experiment 2, none of the two is available and the result suggests that orientation is more dependent on copresence than depicted events are. In Experiment 3, event information has disappeared while orientation information is still present and therefore takes priority over the depicted events. These findings are consistent with the Coordinated Interplay Account outlined by Knoeferle and Crocker (2007). Their account posits that utterance-mediated attention to depicted events mentioned by the verb results in the rapid use of that event information. Additionally, they argue that while memory of prior scene events, including completed actions, can still influence on-line comprehension, information which is no longer visible may decay, resulting in diminished influence. Our findings extend this position, suggesting that not only the visual presence of events matters, but also that of character orientation.

While differences in the time course and the impact of event information due to different presentations are highlighted by our experiments, the use of depicted event information in syntactic disambiguation was shown to generalize to plausible scenes, recently inspected scenes and dynamic events. The robust use of depicted events in on-line disambiguation supports the view that visual scene information plays a fundamental role in situated language comprehension.

Acknowledgments

This research was funded by a PhD scholarship to the first author awarded by the German Research Council.

References

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of memory and language*, *38*, 419–439.

Altmann, G. T. (2004). Language mediated eye-movements in the absence of a visual world: the blank screen paradigm. *Cognition*, *93*(3), 79–87.

Altmann, G. T., & Kamide, Y. (1999). Incremental interpretation at verbs: restricting the domain of subsequent reference. *Cognition*, *73*(3), 247–264.

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

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 (1)*, 37-55.

Knoeferle, P., & Crocker, M. W. (2006). The co-ordinated interplay of visual information and world knowledge. *Cognition*, *30 (3)*, 481-529.

Knoeferle, P., & Crocker, M. W. (2007). The influence of recent events on spoken language comprehension: evidence from eye movements. *Journal of Memory and Language*, *57 (2)*, 519-543.

Knoeferle, P., Crocker, M. W., Scheepers, C., & Pickering, M. J. (2005). The influence of the immediate visual context on incremental thematic role-assignment. *Cognition*, *95*, 95-127.

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

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*(5217), 1632–1634.

Weber, A., Grice, M., & Crocker, M. (2006). The role of prosody in the interpretation of structural ambiguities: a study of anticipatory eye movements. *Cognition*, *99*, B63-B72.