# Visual constraints modulate stereotypical predictability of agents during situated language comprehension

**Alba Rodríguez (arodriguez@cit-ec.uni-bielefeld.de)[1]**
**Michele Burigo (mburigo@cit-ec.uni-bielefeld.de)[1]**
**Pia Knoeferle (pia.knoeferle@hu-berlin.de)[2]**

[1]Cognitive Interaction Technology Excellence Cluster, Bielefeld University, Bielefeld
[2]Department of German Language and Linguistics, Humboldt University, Berlin

**Abstract**

We investigated how constraints imposed by the concurrent visual context modulate the effects of prior gender and action cues as well as of stereotypical knowledge during situated language comprehension. Participants saw videos of female or male hands performing an action and then inspected a display showing the faces of two potential agents (one male and one female face) as they listened to German OVS sentences about stereotypically female or male actions. Unlike previous experiments (Rodriguez et al., 2015), the display concurrent with the sentence also showed a picture of the object of the videotaped action and a 'competitor object' (with opposite stereotypical valence) that had not appeared in the video but could be mentioned in the sentence. We measured eye movements to the faces of the agents during comprehension. The design manipulated the match between the videotaped action and the action described by the sentence (*action-verb(phrase) match*) and the match between the stereotypical valence of the verbally described action and the *gender* of the agent of the previous video (conveyed only by the hands; *stereotypicality match*). We replicated the results obtained in Rodríguez et al. (2015): an overall target agent preference (i.e. the agent whose gender matched that of the hands seen in the previous video), reduced by action-verb mismatches. However, unlike in their study, mismatch effects emerged earlier. In addition, stereotypicality effects emerged in the verb region. The earlier mismatch effects and added stereotypicality effects suggest that the visual availability of the objects, perhaps jointly with the verbal input, facilitated the activation of representations from the recent videos (speeding up mismatch effects) and the consideration of alternative representations, favoring stereotypical expectations.

**Keywords:** Visual constraints; situated language comprehension; gender; stereotypes; eye-tracking.

## Introduction

It is a well-established finding that looks towards objects are closely time-locked to their referring expressions (Tanenhaus, Spivey-Knowlton, Eberhard & Sedivy, 1995). Moreover, comprehenders can even make anticipatory eye-movements to potential referents during incremental sentence comprehension if semantic restrictions or world-knowledge impose a sufficiently constraining context. The incremental comprehension of sentences can activate mental representations based on our experiential, long-term knowledge, and these can guide our attention in a concurrent scene (Altmann & Kamide, 1999; Altmann & Kamide, 2004; Kamide, Altmann & Haywood, 2003). However, when prior visual information also acts as a cue during situated language comprehension (e.g., previously seen agents acting upon patients or objects), comprehenders appear to preferentially rely on these recent actions rather than on their stereotypical thematic role knowledge to predict which entity is mentioned next. If prior visual events are available, comprehenders tend to visually anticipate the characters that had been recently seen as involved in those events, rather than guessing who should have been involved based on other types of knowledge (Knoeferle & Crocker, 2007; Rodríguez, Burigo, & Knoeferle, 2015). Entities from previous events are preferentially looked at during comprehension even when verbal information is at odds with the previously inspected actions (Abashidze, Knoeferle; 2015; Rodríguez, Burigo & Knoeferle, 2015).

In eye-tracking experiments on situated language comprehension, Rodríguez, et al. (2015) contrasted gender and action cues from previously seen events with knowledge of gender stereotypes. Participants inspected videos of female or male hands performing an action. After inspecting the videos, they listened to non-canonical German OVS sentences about stereotypically male or female actions while inspecting another display showing the picture pairs of two potential agents (one male and one female face). The non-canonical, yet grammatical word order permitted us to first mention the event theme (i.e. object) and the verb, and observe participants' anticipation of an upcoming agent by monitoring eye movements to the agents' faces. One of the agents on the display was the 'target agent' (i.e. the agent whose gender matched that of the hands seen in the video), while the other character was the opposite gender competitor. Participants' task was to verify whether the content of the sentence matched the immediately preceding event via a button press. In their first experiment, Rodríguez et al. (2015) manipulated two factors. First, whether the videotaped action matched or mismatched the action described by the sentence, i.e. *action-verb (phrase) match*; and second, whether the gender-stereotype associated with the verbally described action matched or mismatched the *gender* of the agent in the

previous video, only conveyed by their hands (i.e. *stereotypicality match*, e.g., congruous case: female hands performed the action in the video and then the sentence was about a stereotypically female action; incongruous: female hands performed the action and then the sentence was about a stereotypically male action, see Table 1). Experiment 2 was similar to Experiment 1, but instead of a mismatch in action-verb reference, the subject of the sentence either matched or mismatched the gender of the agent in the video. These results showed that visual (gender) cues were easily extracted from prior events (i.e. the pair of hands acting upon an object) and successfully integrated with information from a subsequent scene depicting potential agents. Participants responded correctly on more than 90% of the trials in both experiments and they showed an overall preference for inspecting the target agent vs. the competitor during comprehension, regardless of the stereotypical content of the sentence. This preference for the target agent was however reduced in cases of action-verb mismatches at the verb (Exp1) and final noun (Exp2) regions.

One possible account for these results and those obtained in other studies (Knoeferle & Crocker, 2007; Abashidze, Knoeferle, 2015) might be that a preceding event gains relevance by virtue of being recent in memory (Morrison, Conway & Chein, 2014). In that sense, mental representations of events during language comprehension might be preferentially derived from recent perceptual information, rather than experiential knowledge in long-term memory. The mismatch effects observed in our previous experiments (Rodríguez et al. 2015) suggest that when the event described by the utterance does not match a previously seen event, participants still show a preference for it(s agent) even if reduced, and do not seem to rely on the described events and long-term stereotypical knowledge in directing their attention. Incongruence in our experiments thus only led to a disengagement of attention from the target agent. How could we encourage event representations that rely more on long-term knowledge and less on the immediate context, especially where language is at odds with the recently perceived events?

## Visual constraints on language comprehension

Both a recent visual context and co-present scenes can affect how rapidly listeners understand language and to which extent they can anticipate upcoming referents. On the one hand, prior scenes can permit anticipation of upcoming entities even when the scene is no longer present during language comprehension and when listeners must rely on previous cues (as was the case in Rodríguez et al., 2015) or a mental representation of a scene. Altmann (2004), for example, had participants inspect a visual context showing a man, a woman, a cake, and a newspaper. After participants had inspected the scene, it was replaced with a blank screen. Next a spoken sentence was played ('The boy will eat the cake'). Participants directed their looks in the blank screen to where the cake had been placed while listening to the word 'eat' while hearing 'the man will eat the cake', and this happened as efficiently as with a concurrent visual context.

Much like a recent visual context can constrain anticipation, listeners' language comprehension and visual attention is also influenced by how the concurrent scene is configured (i.e. more or less constrained). One influential study tested how changes in the concurrent visual context affected the resolution of syntactic ambiguity in sentences like *Put the apple on the towel in the box* (Tanenhaus et al. 1995). An example scene contained either just one referent for an apple (i.e. an apple on a towel) and an empty towel, or it contained a second apple (i.e. on a napkin). When only one apple was present ('one referent' condition), participants tended to incorrectly fixate the empty towel as the goal for the apple after hearing the modifier *on the towel*. However, when two apples were present ('two referent' condition) participants rarely looked at the empty towel (the incorrect goal). Thus, the presence of two referents in the scene prompted participants to interpret *on the towel* as the modifier of the noun *apple*, thus resolving the syntactic ambiguity against their preferred analysis (attachment into the verb phrase).

In terms of the configuration of the visual context in the experiments by Rodríguez et al. (2015), the scene might have been too *unconstrained*: The presence of only the pictures of the potential agents may have biased participants to recruit one type of information (i.e. the gender of the agent in the previous video) when verifying the content of the sentence. While the agent could be easily linked to the video, information about the objects (and their corresponding actions) was no longer available in the concurrent scene. Consequently, participants may have been discouraged from exploiting other sources like long-term knowledge of stereotypes during comprehension. The appearance of objects, together with verbally expressed actions (i.e. the verb phrase), may impose additional constraints on participants' representations, both of recent events as well as of new alternative representations derived merely from language (especially when the sentence mismatches the preceding events). At the same time, world-knowledge associated with the representation of actions performed upon those objects could also be enhanced (Blair, & Banaji, 1996; Bargh, 1999), motivating participants' inferences and expectations towards the upcoming agents.

In summary, previous research has shown that listeners' expectations of soon-to-be-mentioned entities during language comprehension can be constrained by multiple sources. The visual context is one of them, sometimes in the form of prior scenes that are no longer present during language comprehension. This is reflected in anticipatory shifts towards the relevant characters or objects, or the locations where such entities had been previously presented. However, manipulating the configuration of the concurrent scene (i.e. which objects it contains) can also impose contextual constraints on real time language processing.

## The current study

The findings from Tanenhaus et al. (1995) and Spivey et al. (2000), which support the idea that manipulating visual constraints can modulate real time language comprehension, motivated a modification in the design employed by Rodríguez et al. (2015, Experiment 1). Participants first inspected a videotaped event that showed a pair of hands interacting with objects (e.g., baking a cake or building a toy model). Recall that in the studies by Rodriguez et al. (2015), the target display following the video only contained photographs of the faces of a male and a female character. By contrast, in the present experiment, we added two pictures: one photograph showed the object that had been acted upon in the preceding video; the other was a photograph of an object that had not been seen before, but that could potentially be mentioned. This latter object was part of an action with opposite stereotype valence from that of the object in the preceding event. For example, after showing a video of a pair of female hands baking a cake, the visual display would contain a female face (the target agent), a male face (the competitor agent), the cake (from the stereotypically female action in the preceding video) and a toy model (a competitor object, part of a model building action which would be stereotypically male, see Figure 1). We measured visual attention to the agent picture pairs during OVS sentence comprehension. Like in our previous experiments, participants answered via button press whether the sentence matched the video they just saw before ("yes" or "no").

The design in the present experiment was the same as in Rodríguez et al. 2015, Experiment 1: action-verb (phrase) match was manipulated together with stereotypicality. Accordingly, we predicted to replicate their findings of shorter response times to action-verb mismatches compared with matches. Additionally, the presence of objects in the target display, together with verbally expressed actions could help boost the representations of action events. If those representations enhanced the activation of gender-related stereotype knowledge, reaction times might be modulated by the stereotypicality of the described actions. This would extend the findings by Rodriguez et al. who failed to observe clear stereotypicality effects in the response times.

For the eye-movements, we initially expected that participants would prefer to inspect the target agent, in line with previous results (Rodriguez et al., 2015) and supporting accounts of visually mediated language comprehension, (Tanenhaus et al., 1995; Knoeferle & Crocker 2007). This preference should decrease when the action described by the sentence mismatches the previously depicted event. However, if changes in the scene that people inspect during comprehension facilitate access to the representation of the recently inspected event as well as the formation of new alternative representations, then action-verb(phrase) mismatch effects could occur earlier than in Rodríguez et al. (e.g., by the end of the first noun or during the verb).

Mismatches in stereotypicality could cause a decrease in preference for the target agent (in line with the gender bias instantiated by the object depictions), particularly for action-verb mismatches. For these, the content of the sentence would not support a representation based on the recently inspected event but rather new representations of action events involving objects that had not been previously seen. The stereotypical beliefs associated with those representations could then be used to anticipate the alternative agent (Blair, & Banaji, 1996; Bargh, 1999). For example, if participants saw female hands backing a cake (a stereotypically female action), but the following sentence described a model building action (stereotypically male), then the presence of an object such as a toy model might bias participants towards looking away from the female face and towards the male agent during sentence comprehension.

# Experiment

## Participants

32 participants took part in the experiment (16 females, 18-32 years). All were German native speakers and had normal or corrected to normal vision. They all gave informed consent before the experiment.

## Materials and Design

We used the materials from our previous studies, which were selected from a norming study: a total of 128 videos of actions (both stereotypically female and male actions were recorded with a female and a male actor; two female actors and two male actors were used). Videos were close-ups of pairs of hands acting upon objects on a table, from external perspective and centered on screen.

From the videos and target displays we constructed 32 experimental items. Each item consisted of a video pair (a pair contained one stereotypically female action, e.g. baking a cake; and one stereotypically male action, e.g. building a model) with their corresponding German sentence pairs with a NP1(object)-V-ADV-NP2(subject) structure. We matched within an item the number of syllables of the words, and the onsets of the different constituents were synchronized. Sentences were presented via speakers, and had a relatively neutral intonation. For the target displays (shown during sentence presentation) we took screenshots of two male actors and two female actors (see Fig. 1) uniformly dressed in black. Additionally for this experiment, the target display also showed snapshots of the objects that belonged to the same item or action pair. One would be the target object (the one that appears in prior events), while the other object would be the competitor object, part of the unseen action with the opposite stereotypical gender valence.

The experimental manipulation followed that of Rodríguez et al. (2015, Experiment 1). We manipulated: a) action-verb congruence (the action described by the sentence matched or mismatched the action previously seen in the video) and b) stereotypicality congruence (the action

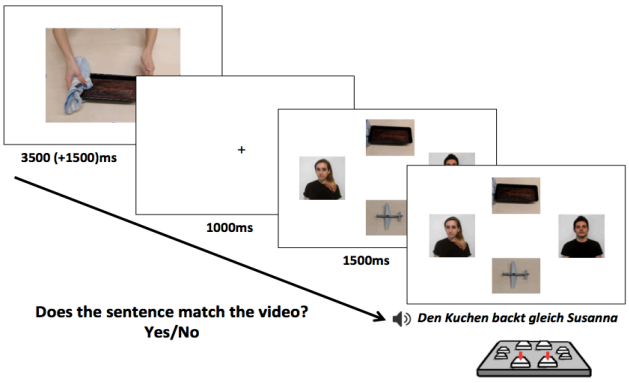


**Figure 1. Example of an experimental trial**.

described by the sentence either matched or mismatched stereotypically with the gender implied by the hands performing the action in the video. The sentence-final subject in the experimental items always matched the agent of the video in terms of gender. The manipulation of the above-mentioned two factors gave rise to four conditions (see Table 1), which were counterbalanced across 8 lists in a Latin Square manner.

Fillers contained videos of actions which were not classified as stereotypically female or male (e.g. filling out a crossword puzzle), with the same sentence structure as the experimental items (18); videos showing two pairs of hands engaged in an action with dative constructions (18: 9 dative-first and 9 dative-middle sentences); and pictures of objects with sentences of different structures (34). Half of the fillers contained video-sentence mismatches of different types (e.g. action, agent gender, and color).

## Procedure

An Eyelink 1000 Desktop Mounted Eye-Tracker recorded participants' eye movements. Viewing was binocular but only the right eye was tracked. Participants completed 10 practice trials before starting the experiment. Trials started with a video of the action for 3500 ms, then the video stopped and the final frame (displaying both the hands in resting position and the object) stayed for another 1500 ms. After that a cross appeared for 1000 ms and then a target screen was shown, with one picture of a female face and another of a male face along the horizontal axis. Along the vertical axis (position was counterbalanced across trials), the target screen showed two objects (one object had featured in the video and the other was a competitor with opposite gender stereotypicality). After a 1500 ms preview, the sentence was presented and eye-movements to the pictures recorded. Participants verified whether the video they just saw matched the sentence that they listened to ("yes" or "no") via button press (Cedrus RB 834). The position of the response buttons was counterbalanced across participants (Figure 1).

## Analysis

Sentences were divided into four time regions, consisting of the first noun phrase region (NP1, object region), the verb region (V), the post-verbal adverb region (ADV) and the final noun phrase region (NP2, subject). Each time region extended from its onset to the onset of the next region except for NP2, which finished at the end of the sentence. For analysis, each time window was shifted forward by 200ms, to account for saccadic planning (Matin, Shao & Boff, 1993; Ferreira et al., 2013).

Because looks to one of the characters implied fewer looks to the other character in the scene, we computed the mean log-gaze probability ratio for each region to measure the bias of inspecting the target agent (i.e. the picture of the character whose gender matched that of the hands in the previous video) over the competitor ($ln(P(\text{target agent})/P(\text{competitor}))$). Values above zero reflect a target agent preference, while values below zero reflect a preference for the competitor. These scores are suitable for parametric tests such ANOVAs (Arai, Van Gompel & Scheepers, 2007; Knoeferle et al., 2011). Mean log probability ratios were calculated (both by subject and by item) for each of the regions separately and then subjected to ANOVAs. Reaction times (RTs) were calculated from sentence onset and accuracy was computed per condition. Missing and incorrect responses were excluded from both eye-tracking and the RT analyses.

## Results

*Accuracy and response times:* Participants responded correctly on 97% of the trials with no significant differences in accuracy between conditions. Response times were faster for action-verb mismatches than for matches ($p < .001$).

*Eye-movement data:* Similarly to previous experiments

**Table 1. Example item with literal translations**

| Video | Sentence | | | | Action-verb match | Stereotypicality match |
|---|---|---|---|---|---|---|
| Female hands baking a cake | ***Den Kuchen***$_{NP1}$ the cake | ***backt***$_{V}$ bakes | ***gleich***$_{ADV}$ soon | ***Susanna***$_{NP2}$ Susanna | Yes | Yes |
| Female hands building a model | ***Das Modell***$_{NP1}$ the model | ***baut***$_{V}$ builds | ***gleich***$_{ADV}$ soon | ***Susanna***$_{NP2}$ Susanna | Yes | No |
| Female hands building a model | ***Den Kuchen***$_{NP1}$ | ***backt***$_{V}$ | ***gleich***$_{ADV}$ | ***Susanna***$_{NP2}$ | No | Yes |
| Female hands baking a cake | ***Das Modell***$_{NP1}$ | ***baut***$_{V}$ | ***gleich***$_{ADV}$ | ***Susanna***$_{NP2}$ | No | No |

(Rodriguez et al., 2015), most mean log ratios remained above zero throughout the sentence, suggesting an overall preference for the target agent. However, an action-verb congruence effect emerged as early as in the NP1 region ($p < .01$, Figure 2A). Participants were more likely to inspect the target agent when the object mentioned in the sentence matched (vs. mismatched) the depicted events. At the verb, main effects of both action-verb ($p = .01$, Fig. 2B) and stereotypicality ($p = .01$) congruence emerged, which prevailed until the ADV region. As during NP1, the target received more looks compared to the competitor in action-verb matches compared to mismatches;; additionally, the target was looked at more when the action described by the sentence was stereotypically congruent in terms of gender. For the final, NP2 region, there was a main effect of action-verb congruence ($p < .01$), and an interaction between verb-action congruence and stereotypicality congruence (in the by subject analysis, $p < .05$).

.

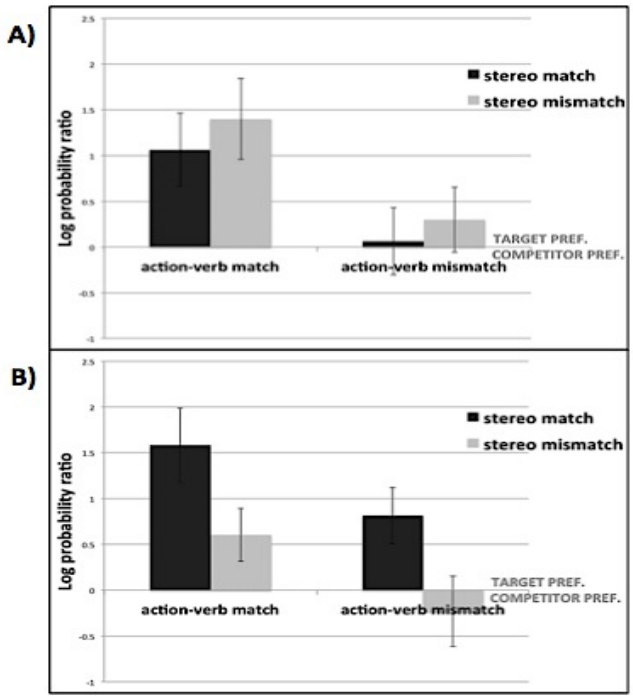


**Figure 2. By-subject mean log probability ratios (SE) in the NP1 region (A) and verb region (B)**

## Discussion

Previous research has contrasted the effects of gender and action cues from a recently inspected event with those of gender stereotype knowledge in situated language comprehension (Rodriguez et al., 2015). These studies have observed no clear effects of gender stereotype knowledge. Instead, analyses of the data from these studies have revealed participants' reliance on gender and action cues derived from the preceding visual context. In the present experiment, we assessed to which extent constraints imposed by the *concurrent* scene might affect the use of recent visual cues as well as boost reliance on knowledge of gender stereotypes. Based on prior research (Rodriguez et al., 2015, Experiment 1), we manipulated the match between the previous event and the sentence (i.e. *action-verb(phrase) match*), as well as between the stereotypicality of the action described and the gender of the previously seen agent (i.e. *stereotypicality match*). Participants saw one video of either female or male hands performing an action and then listened to a German OVS sentence about a stereotypically female or male action, while inspecting a target scene. Unlike in prior research (Rodríguez et al., 2015), participants inspected not just photographs of the faces of two potential agents in the target scene, but also a photograph of an object from the previously seen event and a so-called 'competitor' object (with opposite gender stereotypicality valence). We reasoned that the additional object presentation together with the sentential verb phrase would facilitate the representation of the recently perceived event as well as of another alternative event when language mismatched the prior event (i.e. action-verb mismatches). The object and event representations could enhance (long term) stereotypical knowledge, potentially encouraging expectations of the alternative competitor (Blair, & Banaji, 1996; Bargh, 1999).

The added contextual constraints gave rise to subtly different results compared with the results reported by Rodríguez et al. (2015). We still did not find differences in the behavioral results for the current experiment compared with the first experiment reported by Rodríguez et al. (2015). Moreover, the overall preference for looking at the target agent (vs. the competitor agent) replicated.

Crucially, however, we did observe more rapid effects of action-verb congruence manipulations in the present study (already in the NP1 region). In addition, the present analyses confirmed effects of stereotypicality which had been absent in Rodriguez et al. (2015). These effects emerged as early as in the verb region of the sentence for both action-verb matches and mismatches: participants' preference for inspecting the target agent was more pronounced when the verbally expressed action matched in terms of gender stereotypicality (e.g. when female hands had performed an action, participants preferred to look at the female agent more when the following sentence mentioned a cake baking action compared to a model building action). Also worth mentioning is that the fully mismatching condition (where the action described mismatched previous events and was stereotypically incongruent with the gender cued by the hands did experience a subtle shift of attention towards the competitor agent.

To date, prior visual events have shown a virtually invariant influence on visual attention during situated language comprehension. These events provide detailed information about actions and their associated target objects, as well as individuating information (e.g. about gender of an agent's hands) that can serve to identify associated agents. Upon the encounter of linguistic information, this recent information, rather than world-knowledge derived from language, predominantly guides people's attention over

entities in a referential manner (e.g. a verb phrase can identify the agent that was involved in prior events, and a referent that matches those features will be fixated during the unfolding of the sentence). We have seen this even in cases in which language was at odds with prior events (Abashidze, Knoeferle; 2015; Rodríguez, Burigo & Knoeferle, 2015). However, even in the presence of those recent events, constrains in the concurrent visual input can potentially boost a greater *in situ* visuo-linguistic interaction.

When it is sufficiently constraining, the concurrent visual context seems to facilitate the detection of visuo-linguistic mismatches early during the sentence. Perhaps the concurrent visual context served to boost the representation of the prior event and this resulted in more rapid integration of the prior event representation with representations of the unfolding sentence, eliciting rapid mismatch effects. Additionally, verbal information together with photographs of related objects seems to allow for the consideration of alternative events, enabling further inferences more in line with world-knowledge. Indeed, when language described a different event from the one that participants had just seen, and when world-knowledge derived from the sentence pointed towards a different agent from the one cued by the preceding event, then the impact of the recent events can be greatly diminished.

The results obtained in this experiment support the idea that manipulating visual constraints in situated language comprehension may not only make a difference in terms of how comprehenders resolve structural ambiguity (Tanenhaus et al. 1995; Spivey et al., 2000). Constraining the visual environment via additional (object) referents (as implemented in the present study compared with Experiment 1 by Rodriguez et al., 2015) can boost concurrent visuo-linguistic interactions and facilitate event representations of preceding events, but also new representations of unseen events on site. This facilitation can modulate the time course with which mismatch effects emerge. The presence of additional entities in the concurrent scene during language comprehension seems to also allow for a greater use of inferences based on world-knowledge, which can modulate the extent to which prior perceptual information is used.

## Acknowledgements

This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement n. 316748 and from the Cognitive Interaction Technology Excellence Cluster 277 (DFG).

## References

Abashidze, D., & Knoeferle, P (2015) Do people prefer to inspect the target of a recent action? The case of verb-action mismatches. *Proceedings of the 21st Conference on Architectures and Mechanisms for Language* Valletta, Malta.

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

Altmann, G. T. M. (2004). Language-mediated eye movements in the absence of a visual world: The 'blank' screen paradigm. *Cognition, 93*, 79-87.

Arai, M., Van Gompel, R. P. G., & Scheepers, C. (2007). Priming ditransitive structures in comprehension. *Cognitive Psychology, 54,* 218-250.

Bargh, J. A. (1999). The cognitive monster: The case against the controllability of automatic stereotype effects. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology*, 361-382.

Blair, I. V., & Banaji, M. (1996). Automatic and controlled processes in stereotype priming. *Journal of Personality and Social Psychology, 70*, 1142-1163.

Ferreira, F., Foucart, A., & Engelhardt, P. E. (2013). Language processing in the visual world: Effects of preview, visual complexity, and prediction. *Journal of Memory and Language L, 69*, 165-182.

Kamide, Y., Altmann, G. T. M. & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language, 49,*133-159.

Knoeferle, P. & Crocker, M. W. (2007). The influence of recent scene events on spoken comprehension: Evidence from eye movements. *Journal of Memory and Language, 57*(4): 519-543.

Knoeferle P., Carminati, M.N., Abashidze D. & Essig K. (2011). Preferential inspection of recent real-world events over future events: evidence from eye tracking during spoken sentence comprehension. *Frontiers in Psychology, 2*:306, 1-12.

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

Morrison, A.B., Conway, A.R.A. & Chein, J.M. (2014) Primacy and recency effects as indices of the focus of attention. *Frontiers in Human Neuroscience*, *8*:6.

Rodríguez, A., Burigo, M., & Knoeferle, P. (2015). Visual gender cues elicit agent expectations: different mismatches in situated language comprehension. *Proceedings of the EuroAsianPacific Joint Conference on Cognitive Science* , 234–239.

Spivey, M. J., Tanenhaus, M. K., Eberhard, K. E., & Sedivy, J. C. (2000). Eye movements and spoken language comprehension: Effects of visual context on syntactic ambiguity resolution. *Cognitive Psychology*, 45, 447-81.

Tanenhaus, M., Spivey-Knowlton, M., Eberhard, K., & Sedivy, J. (1995). Integration of visual and linguistic information during spoken language comprehension. *Science, 268*, 1632-1634.