Anticipation in real-world scenes: The role of visual context and visual memory

Moreno I. Coco (University of Edinburgh), George L. Malcolm (University of Glasgow), & Frank Keller (University of Edinburgh)

mcoco@staffmail.ed.ac.uk

Situated language comprehension; Anticipation; Visual context; Visual memory

When we comprehend sentences in the context of visual scenes, we generate expectations about upcoming linguistic material (Altmann and Kamide, 1998). These expectations are constrained by the current input and are incrementally revised (Knoeferle and Crocker, 2006). So, upon hearing *the man will eat the sandwich*, while the verb *eat* unfolds, anticipatory eye-movements are launched to the semantically appropriate object SANDWICH, if visually depicted.

Contextual expectations of a similar nature contribute to visual guidance during visual search in real-world scenes: when looking for MUGS, fixations are more likely on COUNTERS than on FLOORS (Torralba, et al., 2006). We therefore hypothesize that expectations extracted from the visual scene also play a role also during situated language processing. So, upon hearing the verb *eat*, we should observe anticipatory eye-movements to the object TABLE, which is contextually related to the action of eating. This should happen even if the search target (SANDWICH)is not depicted.

We test this hypothesis in Experiment 1: participants listened to sentences (e.g., *the man ate the sandwich*), while viewing scenes containing a target object (SANDWICH) and a contextually related object (TABLE), among other objects. In order to maximize contextual expectations, we used photo-realistic scenes, as these are maximally contextually coherent. In a 2x2 design, we manipulated the thematic restrictions of the verb (specific: *eat*; ambiguous: *move*) and the presence of the target object (present, absent). In a linear mixed effect analysis of the time-course of fixations from verb onset to 1000ms after it, we observed anticipatory looks to the contextually related object TABLE driven by the thematic restrictions of the verb. Crucially, this effect was found regardless of whether the target object was visually present. This demonstrates that (a) anticipatory eye-movements are generated in complex real-world scenes, not just in visual arrays or clip-art scenes (used in the prior literature); and (b) the semantic context provided by a scene can constrain the incremental interpretation of situated speech. This constraint is so strong that it operates even in the absence of the target object.

Experiment 2 aimed to establish whether the effect found in Experiment 1 is a memory effect (possible target locations are stored in memory, as in visual search, Torralba et al., 2006), or whether contextual expectations are computed on the fly, requiring the co-presence of visual and linguistic information. We used the blank-screen paradigm, which previously demonstrated anticipatory eye-movements driven by the thematic restrictions of the verb, even when the scene was no longer present (Altmann, 2004). The experimental conditions and materials were the same as in Experiment 1, but participants previewed the scene for 5000ms before it disappeared. Then after a 1000 ms pause the sentence was played.

In this setting, we failed to find anticipatory effects; the target region on the blank screen was fixated only once the post verbal NP was processed and only when the object had been depicted during preview. This suggests that (a) the blank-screen anticipation effect found by Altmann (2004) does not generalize to real-world scenes; and (b) contextual expectations are computed on the fly, i.e., they require the scene and the linguistic input to be co-present. The memory trace of a visual scene is sufficient to locate the target object, but it cannot be used to infer where the target object should have been, given the memorized visual context: (i.e., a previously seen TABLE is not enough to infer that a SANDWICH could have been there).

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.

Knoeferle, P., & Crocker, M. W. (2006). The coordinated interplay of scene, utterance and world knowledge. Cognitive Science, 30, 481–529.

Torralba, A., Oliva, A., Castelhano, M., & Henderson, J. (2006). Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. Psychological review, 4(113), 766–786.