UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

Rapid extraction of event participants in caused motion events.

Permalink
https://escholarship.org/uc/item/1z6685vb

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 33(33)

ISSN
1069-7977

Authors
Wilson, Frances
Papafraou, Anna
Bunger, Ann
et al.

Publication Date
2011

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Rapid Extraction of Event Participants in Caused Motion Events

Frances Wilson (fwilson@psych.udel.edu)
Department of Psychology, 108 Wolf Hall, University of Delaware, Newark, DE 19716

Anna Papafragou (papafragou@psych.udel.edu)
Department of Psychology, 108 Wolf Hall, University of Delaware, Newark, DE 19716

Ann Bunger (bunger@udel.edu)
Department of Psychology, 108 Wolf Hall, University of Delaware, Newark, DE 19716

John Trueswell (trueswel@psych.upenn.edu)
Department of Psychology, University of Pennsylvania, 3401 Walnut Street, Philadelphia, PA 19104

Abstract

When viewing a complex event, it is necessary to identify and calculate the relationships between different entities in the event. For example, when viewing a caused motion event (e.g. a man raking leaves into a basket.), people need to identify the Agent (man), the affected object or Patient (leaves), the Instrument (rake) and the Goal (basket). In this paper we explore how this process of event apprehension proceeds using eye-tracking methodology. Our study indicates that viewers extract event components rapidly, but some components can be extracted faster than others. Moreover, there is a structure to saccade patterns when participants are asked to identify specific event components. In caused motion events, attention is allocated to the Patient during the early stages of processing even when the Patient is not the target. We discuss implications of this work for how people perceive complex events.

Keywords: spatial cognition; event perception; thematic roles; eye-tracking

Introduction

Perceiving and understanding events in the world is an important part of human cognition. Visual input is highly complex, and yet people are able to rapidly extract information about the basic level category of a scene (e.g. a highway scene) as well as objects within a scene (e.g. Biederman, 1995; Oliva & Torralba, 2001; Oliva, Torralba, Castelhano & Henderson, 2003; Potter, 1975). In addition, when we view a scene or event we need to determine the relations that exist between different elements in the scene or different event participants. For instance, when we see a man hitting a ball, we need to conceptualize the causer of the event (or Agent—here the man) and the entity directly affected by the action (the Patient—here the ball). More complex representations of the event may include the Instrument used for hitting and the Goal or destination of the moving ball. Identifying both the types of event components that viewers are able to extract from dynamic events and the time course of extraction of individual event components is important for understanding how people process visual information. Additionally, since these event roles correspond fairly straightforwardly to linguistic information ("thematic roles;" see Dowty, 1991; Koenig Mauner & Bienvenue, 2003; Henderson & Ferreira, 2004), the processes underlying non-linguistic event apprehension can be informative about theories of how people produce, understand and acquire language. However, the field has only begun to investigate the question of how humans succeed in parsing ongoing events.

In an early study of event apprehension, Griffin and Bock (2000) examined eye-movements to a still image depicting an event with two animate participants (e.g. a woman shooting a man). When participants freely inspected the image, they showed a preference for fixating Patients over Agents after 1300ms of inspection. But when participants were instructed to find the event Patient, fixations to the Agent and Patient began to diverge early, after approximately 300ms. These findings suggest that Patients can be identified rapidly, and are allocated attention after initial scene processing.

Webb, Knott and MacAskill (2010) extended Griffin and Bock's study to "reach-to-grasp" actions, using video presentation of a human agent reaching to grasp an inanimate object (e.g. a green building block). Unlike Griffin and Bock, they found that participants made early fixations to human agents. However, as the authors acknowledge, it is not clear whether these findings show an early preference for looks to Agents or simply to moving, animate entities, which are known to attract attention (Abrams & Christ, 2003). Despite this limitation, this study does show a temporal structure in attention to event components, with attention starting with the origin of the action (the Agent), then moving to the anticipated location of the Patient. However, it is not clear whether this finding would generalize to events where there are no disparities of animacy and motion.

Using a rapid presentation of scenes, Dobel, Gumnior, Bölte and Zwitserlood (2007) showed that information about the relationships between event components can be extracted rapidly. In scenes that depicted an Agent, a Patient and a Goal/Recipient (such as an archer shooting an arrow to a target), judgments about the coherence of the scene were made accurately even at very short presentations of, e.g. 100ms. At presentation durations of 250-300ms the Agents were named more accurately than the Goals (approx. 75% vs. 60%), again suggesting that Agents may be privileged over other event components. Patients were named less accurately, but it is possible that this was due to the relatively small size of the Patients relative to the other event components. Dobel and colleagues concluded that such rapid apprehension of scene coherence suggests that roles within an event can be assigned without fixation on the relevant area of the scene. However, since the decision about scene coherence was made after stimulus presentation, it is possible that subsequent processing based on the representation of the scene in visual memory allowed accurate judgements, rather than processing during stimulus presentation.

Current Study

Here we report an eye-tracking experiment that examines the relation between event components, and the role they play in building a representation of an event. Unlike prior studies that have used relatively simple events, often with only an Agent and a Patient, our study focuses on caused motion events in which an animate Agent uses a tool or body part (Instrument) to move an inanimate object (Patient) towards an inanimate target or destination (Goal). We adapted Griffin and Bock's "Find the Patient" paradigm and asked viewers to rapidly identify and fixate each of the four event components present in the event. By examining the speed at which event components can be identified and the pattern of fixations made before fixating the target object, we hoped to determine the relationship between individual event components as event representations are assembled.

We were particularly interested in comparing event role apprehension for the three non-Agent roles (Patients, Goals and Instruments). (Agents in our study were always animate and therefore conflated animacy and agency.) Our study asked whether these event components can be identified by viewers equally rapidly and/or independently from one another. There are at least two possibilities about how such event roles are extracted from caused motion sequences. According to Dobel and colleagues (2007), information about event roles can be extracted in the earliest stages of scene presentation. If this is the case, then we would expect participants to saccade directly to the target event component, with no systematic pattern of prior fixations. Another possibility is that, even if extraction of event components is generally rapid, it might not be equally rapid for all event components. On the basis of Griffin and Bock's (2000) data, which found that attention is directed towards the Patient more than the Agent, we might expect the Patient role to be easier to identify than other components, and see early fixations on the Patient in all conditions. This possibility is supported by evidence from other domains. Linguistic evidence suggests that different event components are not accorded the same status (Koenig, Mauner & Bienvenue, 2003). Verb arguments are typically considered to be part of the lexical entry for a verb, and thus obligatory, while adjuncts are optional. Boland and Boehm-Jernigan (1998) provide evidence that arguments are read faster than adjuncts, suggesting that arguments and adjuncts are distinguished by the sentence processor. Agents and Patients are usually encoded as verb arguments, while Instruments are typically accorded adjunct status (Boland, 2005). The status of Goals with respect to the argument/adjunct distinction is less clear: while they are required by the subcategorization frames of certain verbs (e.g. *put*), they show variability with respect to the preposition used, in contrast to the prototypical prepositional argument taken by dative verbs (e.g. *show this to Simon*) (Tutunjian & Boland, 2008). If the non-linguistic processing of event components reflects the way in which they are encoded linguistically, then we might expect that Patients are identified more easily than Goals, and Goals more easily than Instruments.

Method

Participants

Forty undergraduate students from the University of Delaware participated for class credit (Mean age =19;1).

Materials

Eighteen test pictures were created using clip art images. The pictures depicted caused motion events, such as a man using a rake to rake leaves into a basket (e.g. Fig. 1). The Agent of each action was always an adult human, and the pictures always included an object affected by the action (the Patient) (e.g. the leaves) and a Goal or destination for the action (e.g. the basket). The Instrument used to perform the action was either a tool (such as a rake) or a body part (such as a foot used for kicking).

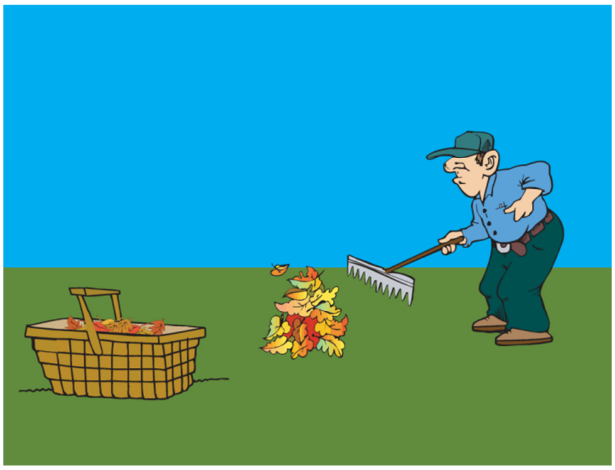

Figure 1 Example Test Item.

An additional set of 18 caused motion events were used as fillers. Filler items alternated with experimental items. Two pictures depicting frogs were created for display after each experimental item and filler to encourage participants to make eye-movements around the screen. Participants were randomly assigned to one of two orders of the stimuli, one the reverse of the other.

Procedure

Participants were told that they would see pictures depicting an action or event. Each participant was randomly assigned to one of four conditions. In the Agent condition, participants were told to look as quickly as possible at "the person or animal who was performing the action," and to press the space bar as soon as they were doing so. In the Instrument condition, participants were given the same instruction but told to look at "the tool or body part used to make the action." In the Goal condition, participants were told to look at "the goal or destination of the action," and in the Patient condition, participants were instructed to look at "the object directly affected by the action." Every participant saw a practice picture (an archer firing an arrow at a target) in which the target item relevant to their condition was highlighted. Before each of the 36 pictures (18 experimental items and 18 fillers), participants were instructed to fixate a cross located at the top of the screen in the center, and to press the space bar when they were fixating it. After each picture, participants viewed one of two pictures (randomly selected) depicting two frogs for 3000ms. Participants' eye-movements were tracked using a Tobii T60 eye-tracker. At the start of the experiment, participants' eye-movements were calibrated using a five-point calibration procedure, in which they followed a red dot which moved to the four corners of the screen and then to the center of the screen. If calibration was incomplete, the procedure was repeated. Typically participants required only one calibration. Participants were seated approximated 60cm from the screen. The experiment took approximately 5-10 minutes.

Results

Coding

In each scene, four Areas of Interest (AOIs) were defined (Agent, Patient, Instrument, Goal) using the Tobii Studio AOI tool. AOIs did not overlap. In cases where the Agent was holding an Instrument, the Agent AOI was defined as the area of the Agent's torso and head, and the Instrument as the tool or Instrument itself, as well as the hand and wrist of the Agent. Trials with greater than 30% trackloss were excluded from the analysis (approx. 1.3%)

Analysis

Figures 2-5 show the proportion of fixations to each event component in each condition. In the Agent condition (Fig. 2), we see early looks to the Agent (at around 120ms) and little consideration of other event components. In the Patient condition (Fig. 3), looks to the Patient diverge early (at around 150ms) from looks to the Goal and Instrument, and later (at around 250ms) they diverge from looks to the Agent. In the Goal condition (Fig. 4), looks to the Goal diverge at around 300ms. In the Instrument condition (Fig. 5), we see an early peak of looks to the Patient before looks to the Instrument diverge (at around 250ms).

To assess the reliability of these findings, we calculated the proportion of looks to each event component during four 200ms time windows, starting from the onset of the stimulus. Because proportion data can sometimes violate assumptions of linear statistical models, we first transformed the proportion data to elogit values following a procedure outlined in Barr (2008). The elogit data were then analyzed using multi-level linear modelling with crossed random intercepts for subjects and items (see Baayen, Davidson and Bates, 2008 for discussion). The model contained a single fixed effect of Condition with four levels (Agent, Goal, Instrument and Patient search). The dependent variable was elogit looking time to the target[1] (i.e. Agent in the Agent condition, Patient in the Patient condition, etc.). The lme4 package in the statistical package R, which we used to conduct the analyses, shows the estimates for each level of the fixed factor relative to a base level and provides comparisons of each level of the factor to the base level. For example, using Agent as the base level, the model would give us the comparison between the Agent and the Goal, the Agent and the

---

[1] Since looks to event components within a condition are negatively correlated, and thus not independent, we compared looks to specific event components across conditions.

Instrument, and the Agent and the Patient. However, we were also interested in contrasts between the other levels of the Condition, e.g. between the Instrument and the Goal. To obtain these contrasts we changed the base level of the model. For example, changing the base level to the Goal, we obtained the contrast between the Goal and the Instrument, the Goal and the Patient, and the Goal and the Agent. By rotating the base level to each of the four levels of the factor Condition we were able to obtain all possible contrasts.

In time window 1 (0-200ms), there were more looks to the target in the Agent condition than in the Instrument condition (t=-2.339, p<0.05), but no other significant differences between looks to the target in the other conditions. In time window 2 (200-400ms), there were more looks to the target in the Agent condition than in all other conditions (Instrument: t=-2.131, p<0.05, Goal: t=3.585, p<0.05, Patient: t=-2.131, p<0.05). After rotating the base level, both the Patient (t=-3.013, p<0.05) and Goal (t=-2.151) conditions showed more looks to the target than the Instrument condition during time window 2 (p<0.05). Together, these results suggest that successfully finding an Agent occurred more quickly than finding any of the other event components; furthermore, finding a Goal or a Patient occurred more quickly than finding an Instrument. In the third (400-600ms) and fourth (600-800ms) time windows, there were no significant differences between conditions. Overall, these results

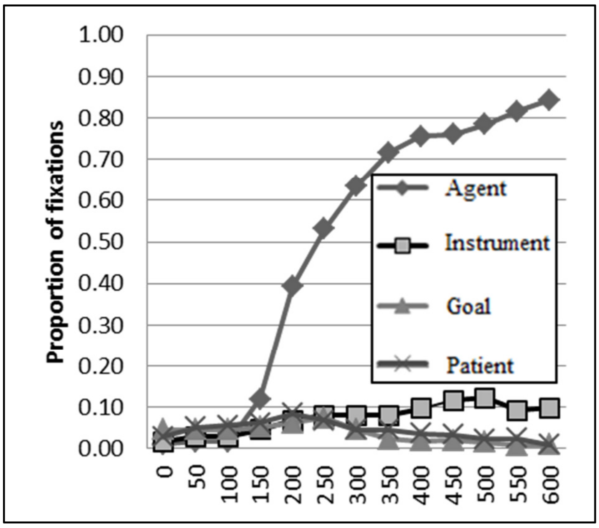

Figure 2 Looks to event components in the Agent condition.

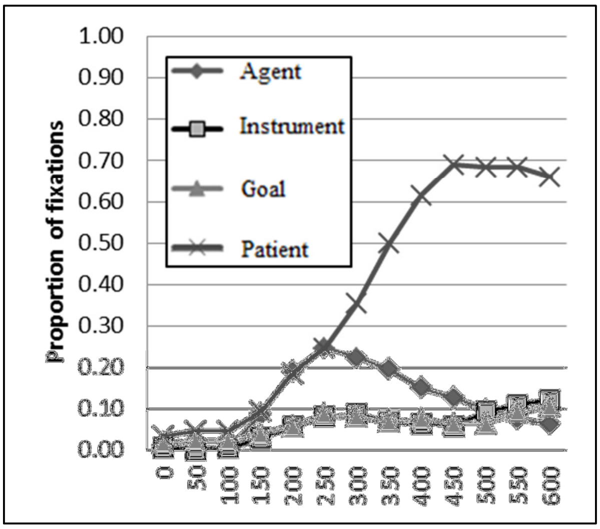

Figure 3: Looks to event components in the Patient condition.

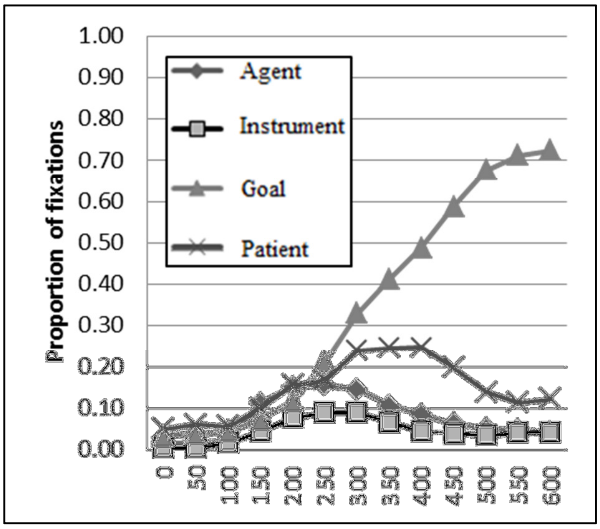

Figure 4: Looks to event components in the Goal condition.

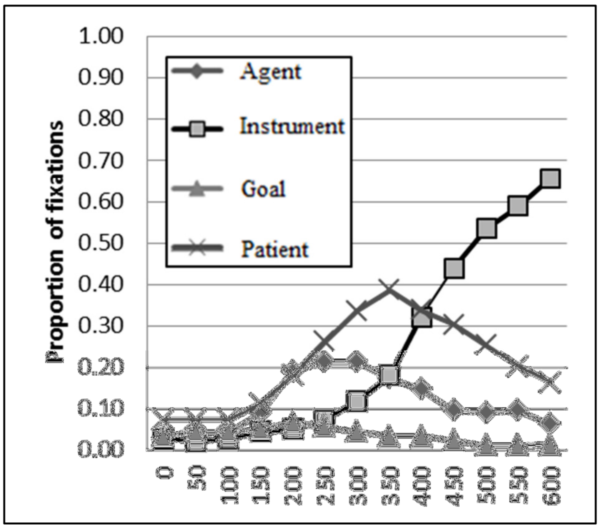

Figure 5: Looks to event components in the Instrument condition.

suggest that event components can be identified rapidly, but point to asymmetries among different event roles.

Could these asymmetries be due to differences in size between AOIs corresponding to individual event components? To preserve scene plausibility, size of event components in our stimuli was not controlled for, and overall, Goals were larger than Agents, which were larger than both Instruments and Patients: paired t-tests confirmed that there were significant differences in size (as measured as a percentage of image area using the Tobii Studio AOI tool) between Goal and Patient (t(17)=-6.77, p<0.0001), Goal and Instrument (t(17)=-6.14, p<0.0001), Goal and Agent (t(17)=-5.05, p<0.0001), Instrument and Agent (t(17)=-2.160, p<0.05) and Patient and Agent (t(17)=--3.41, p<0.01). Crucially, however, there was no significant difference in size between the Patient and Instrument (3.5% vs 3.6% of image area), so it does not seem likely that the difference in speed of identification between Patients and Instruments is due to differences in area. Furthermore, although Goal AOIs were bigger, on average, than Patient AOIs, we do not see a difference in speed of identification between Goals and Patients. Finally, and most importantly, the time taken to fixate the Goal, Instrument or Agent did not correlate with AOI size. Only in the Patient condition was a significant negative correlation observed (r=-.508, n=18, p<0.031), indicating that smaller Patients were fixated slightly later. We discuss alternative explanations for the differences between identification of event components in the general Discussion below.

Our initial analysis indicated that the conditions differed most in time window 2. To determine whether there were early fixations on individual event components, we compared looks to each of the event components across conditions in time window 2, using the same model selection procedure as in the previous analysis. For Agents, Goals and Instruments, we found little variation in looks to the relevant component across conditions in which that component was not the target: Starting with Agents, looks to the Agent differed only between the Agent condition and each of the other conditions (Goal, t=-10.551, Instrument, t=-9.113, Patient, t=-8.956, all p<0.05). Similarly, looks to the Goal differed only between the Goal condition and each of the other conditions (Agent, t=-3.832, Instrument, t=-3.982, Patient, t=-3.755, all p<0.05). Finally, looks to the Instrument differed only between the Instrument condition and each of the other conditions (Agent, t=-2.098, Goal, t=-2.144, Patient, t=-2.281, all p<0.05). However, there were more looks to the Patient in the Goal (t=4.669, p<0.05), Instrument (t=7.327, p<0.05) and Patient (t=9.393, p<0.05) conditions than in the Agent condition. After rotating the base level, we found that there were more looks to the Patient in the Instrument condition compared to the Goal condition (t= 2.672, p<0.05). One possibility is that the increased looks to the Patient in the Instrument condition are due to the relatively small sizes of each of these event components (3.5% and 3.6% of image area), which might have led participants to look around the scene to find the target. However, if this were the case, then we would expect to see more looks to the Instrument in the Patient condition, which we do not. Additionally, we would not expect to see additional looks to the Patient in the Goal condition. We consider plausible explanations of such looks to the Patient below.

Discussion

This study sought to investigate the processing of event components in a “Find the X” task. In contrast to previous work in this area, which has mostly investigated the relation between Agents and Patients, we advanced the empirical domain of inquiry by examining the relations between Patients, Goals and Instruments. Our study reveals three major conclusions. Firstly, consistent with the findings of Dobel et al. (2007), we observed that event components could be identified rapidly and accurately (although participants were only able to saccade directly to the target in the Agent condition, where it is probable that participants were relying on animacy cues). Secondly, we discovered asymmetries between event components: not all event components were identified with equal speed. Consistent with Griffin and Bock (2000), we found that Patients were identified particularly rapidly. Furthermore, we found that Instruments were identified more slowly than either Patients and Goals. Our data support the hypothesis that roles typically encoded as arguments in language (e.g. Patients) are identified more quickly than those typically identified as adjuncts (e.g. Instruments). The fact that Goals are identified just as quickly as Patients even though Goals are not prototypical arguments may be related to the high salience of Goals (Lakusta & Landau, 2005; Papafragou, 2010). Although it is not possible to draw firm conclusions about the relation between linguistic encoding and event components at this stage, our data raise the possibility that the distinction in language between arguments and adjuncts is a result of prioritization of event components in non-linguistic processing.

A third, more tentative conclusion can be drawn regarding the role of Patients. The analysis of looks to the Patient component across conditions highlighted an asymmetry between Patients and other event components. While there were no differences in looks to the Agent, Goal and Instrument in conditions in which each component was not the target, looks to the Patient varied across conditions. Unsurprisingly there were more looks to the Patient in the Patient condition (i.e. when it was the target), but somewhat surprisingly, there were more looks to the Patient in the Goal condition compared to the Agent condition, and in the Instrument condition compared to the Goal condition. This result suggests that attention is allocated to the Patient even when the Instrument is the target. Why might this be so? One possibility is that the Patient is somehow more central to the event, and that identifying what has been affected by the action facilitates location of the Instrument. Furthermore, since the Patient is depicted as moving towards the Goal, allocation of attention towards the Patient might facilitate calculation of the trajectory towards the Goal and identification of the location of the Goal within the scene. Alternatively, increased looks to the Patient could be considered further evidence for the distinction between arguments and adjuncts: participants may allocate attention to event components which are typically encoded as arguments (such as Patients) before allocating attention to less prototypical arguments (Goals) and adjuncts (Instruments). However, at this stage it is impossible to draw firm conclusions about the precise nature of the role of Patients in the identification of other event components.

To summarize, we have shown that event components can be rapidly and accurately identified in a scene. However, different event components (Patient, Goal, Instrument) are not identified equally quickly, in a way that may be consistent with the linguistic distinction between arguments and adjuncts.

Acknowledgments

This research was partly supported by NIH/NICHD Grant 3R01HD055498 to A.P. and J.T. Thanks to James Delorme for assistance in data collection and Rick Chalton for assistance in preparation of stimuli.

References

Abrams, R., & Christ, S. (2003). Motion onset captures attention. *Psychological Science*, 14, 427–432.

Barr, D. J. (2008). Analyzing ”visual world” eyetracking data using multilevel logistic regression. *Journal of Memory and Language,* 59(4), 457–474.

Baayen, R. H., Davidson, D. J. & Bates, D. M. (2008), Mixed-effects modelling with crossed random effects for subjects and items. *Journal of Memory and Language* ,59(4), 390–412.

Biederman, I. (1995). Visual object recognition. In M. Kosslyn & D. N. Osherson, eds., *An Invitation to Cognitive Science: Visual Cognition* (2nd edition), vol. 2.

Boland, J. E. (2005). Visual arguments. *Cognition,* 95, 237–274.

Boland, J., & Boehm-Jernigan, H. (1998). Lexical attachments and prepositional phrase attachment. *Journal of Memory and Language*, 29, 684–719.

Dobel, C., Gumnior, H., Bölte, J., & Zwitserlood, P. (2007). Describing scenes hardly seen. *Acta Psychologica*, 12, 129–143.

Dowty, D. (1991). Thematic proto-roles and argument selection. *Language*, 67, 547–619.

Griffin, Z., & Bock, K. (2000). What the eyes say about speaking. *Psychological Science*, 11, 274–279

Henderson, J., & Ferreira, F. (eds.) (2004). *The interface between language, vision and action: Eye movements and the visual world*. New York: Psychology Press.

Koenig, J.-P., Mauner, G., & Bienvenue. (2003). Arguments for adjuncts. *Cognition*, 89, 67–103.

Lakusta, L., & Landau, B. (2005). Starting at the end: The importance of goals in spatial language. *Cognition*, 96, 1–33.

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: a holistic representation of the spatial envelope. *International Journal in Computer Vision*, 42, 145–175.

Oliva, A., Torralba, A., Castelhano, M. S., & Henderson, J. M. (2003). Top-down control of visual attention in object detection. *Proceedings of the IEEE International Conference on Image Processing*, vol. I, 253–256.

Papafragou, A. (2010). Source-goal asymmetries in motion representation: Implications for language production and comprehension. *Cognitive Science*, 34, 1064–1092.

Potter, M. C. (1975). Meaning in visual search. Science, 187, 965–966.

Tutunjian, D., & Boland, J. E. (2008). Do we need a distinction between arguments and adjuncts? Evidence from psycholinguistic studies of comprehension. *Language and Linguistics Compass*, 2, 641–646.

Webb, A., Knott, A., & MacAskill, M. R. (2010). Eye movements during transtive action observation have sequential structure. *Acta Psychologica*, 133, 51–56.