Cognition 206 (2021) 104516

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Short Communication

Speaking for seeing: Sentence structure guides visual event apprehension

Sebastian Sauppe [a,b,*], Monique Flecken [c,d]

[a] Department of Comparative Language Science, University of Zurich, Switzerland
[b] Interdisciplinary Center for the Study of Language Evolution, University of Zurich, Switzerland
[c] Department of Linguistics, University of Amsterdam, The Netherlands
[d] Neurobiology of Language Department, Max Planck Institute for Psycholinguistics, The Netherlands

ARTICLE INFO

Keywords:
Event cognition
Visual attention
Scene apprehension
Sentence production
Brief exposure
Syntax

ABSTRACT

Human experience and communication are centred on events, and event apprehension is a rapid process that draws on the visual perception and immediate categorization of event roles ("who does what to whom"). We demonstrate a role for syntactic structure in visual information uptake for event apprehension. An event structure foregrounding either the agent or patient was activated during speaking, transiently modulating the apprehension of subsequently viewed unrelated events. Speakers of Dutch described pictures with actives and passives (agent and patient foregrounding, respectively). First fixations on pictures of unrelated events that were briefly presented (for 300 ms) next were influenced by the active or passive structure of the previously produced sentence. Going beyond the study of how single words cue object perception, we show that sentence structure guides the viewpoint taken during rapid event apprehension.

1. Introduction

Perception is not a process solely driven by bottom-up input. To the contrary, it is strongly guided by top-down factors, related to perceivers' prior expectations, knowledge, the current context, and task goals (e.g., Gilbert & Li, 2013; Lupyan, Abdel Rahman, Boroditsky, & Clark, 2020; Summerfield & de Lange, 2014). This holds for the processing of basic percepts, such as the orientation, size and identity of single objects (Summerfield et al., 2006), but also for more complex scenes. Already early stages of visual processing such as the rapid extraction of the "gist of a scene" (Castelhano & Henderson, 2007; Henderson & Ferreira, 2004) are conceptually guided (e.g., Henderson, Brockmole, Castelhano, & Mack, 2007). For example, people's prior experiences can enhance the detection of objects or basic scene category information (Biederman, 1981; Hollingworth & Henderson, 1998; Potter & Levy, 1969; Schyns & Oliva, 1994).

For object perception, language can provide rapid online conceptual guidance (Lupyan, 2012; Lupyan et al., 2020): Linguistic labels provide effective cues to perception because the conceptual representation evoked by a label includes a category-diagnostic sensory representation of the concept, so that hearing or reading the word "dog" activates a visual image of a dog. Activating this sensory representation prior to receiving actual perceptual input attunes the visual system to the expected percept and provides top-down feedback during stimulus processing, also when the to-be-perceived object is masked or degraded (Boutonnet & Lupyan, 2015; Lupyan & Ward, 2013; Ostarek & Huettig, 2017; Samaha, Boutonnet, Postle, & Lupyan, 2018). Linguistic labels thus cause, in Lupyan's (2012) terms, *temporary perceptual warping*.

However, the previous focus on single words leaves two knowledge gaps. First, is the perception of complex visual scenes (with relational structure) also susceptible to cueing effects by language? Second, can the syntactic structure of entire sentences (and their underlying conceptual structure) guide initial scene processing? Moving beyond single words and objects is a crucial step forward in unraveling how language interacts with vision, since objects are often observed in a relational context and we typically speak in sentences, not just single words. Of specific interest for addressing these issues are depictions of *events*—dynamic activities happening across time and space (e.g., someone cutting an apple). Central to understanding events are the relations between the participants involved in them, in terms of their *event roles* (Rissman & Majid, 2019; Zacks, 2020). Agents (the "doers"), patients (the "undergoers") and their relation (defining the event type, e.g., dressing or cutting) comprise the abstract, hierarchical structure of an event (Cohn & Paczynski, 2013; Jackendoff, 1990). These event role configurations are conceptual in nature, as they are not dependent on specific realizations of roles and their relations (e.g., Dowty, 1991;

* Corresponding author at: Department of Comparative Language Science, University of Zurich, Switzerland.
E-mail address: sebastian.sauppe@uzh.ch (S. Sauppe).

https://doi.org/10.1016/j.cognition.2020.104516
Received 25 June 2020; Received in revised form 5 November 2020; Accepted 11 November 2020
Available online 20 November 2020

Rissman & Majid, 2019). This information can be extracted from visual stimuli effortlessly, even under very short viewing conditions (less than 100 ms: Dobel, Gumnior, Bölte, & Zwitserlood, 2007; Glanemann, Zwitserlood, Bölte, & Dobel, 2016; Hafri, Papafragou, & Trueswell, 2013; Hafri, Trueswell, & Strickland, 2018) and from early on in infancy (Galazka & Nyström, 2016; Johnson, 2003; Spelke & Kinzler, 2007). Early-stage visual event processing is immediately geared towards the extraction of conceptual and relational information on event roles and types. The ability to extract conceptual event structures rapidly suggests that events are critical units of representation in cognition (Richmond & Zacks, 2017; Zacks, 2020).

Events are also central to communication: We often talk about the events happening around us. When describing an event, one needs to package its conceptual structure into a sentence. This entails linearizing the linguistic expression of event roles and expressing a viewpoint on the event, during the construction of the sentence's message (the process of *perspective taking*, Bock, Irwin, & Davidson, 2004; Levelt, 1989, 1999). For example, the event of a woman dressing a man (cf. Fig. 1) can be expressed with an active ("The woman is dressing the man") or a passive sentence ("The man is being dressed by the woman") in many languages. The core event structure, in terms of who is doing what to whom, expressed by these two sentences is the same: the woman is the agent and the man is the patient, and the relation between them involves some form of physical contact and transfer. Active and passive sentences differ, however, in the viewpoint selected by the speaker. While actives foreground the agent, passives put the patient in the foreground and the agent in the background in the conceptual structure of the event (Bock et al., 2004; Kazenin, 2001; Keenan & Dryer, 2007).[1] The backgrounding can be so strong that the agent can even be left unmentioned in passive sentences ("The man is being dressed"). The conceptual backgrounding of agents in passives is also shown in experimental work: For example, speakers of English were more likely to produce passives when describing stimuli in which the agent was visually less prominent (i.e., when only the agent's hands were shown, and not their face and torso, (Rissman, Woodward, & Goldin-Meadow, 2019). When the agent was thus backgrounded perceptually, speakers foregrounded the patient linguistically. Further, during event description, German speakers also placed fewer fixations on agents, and more fixations on patients, when planning passives as compared to actives (Sauppe, 2017b).

Can event viewpoints as conveyed by different syntactic structures guide information uptake during the rapid apprehension of upcoming scenes? More specifically, can the production of active and passive sentences, and their underlying conceptual structure bias visual attention to events in subsequently presented visual stimuli, analogous to single labels cueing object perception? Such attentional bias should arise through the pre-activation of an abstract event structure by the syntactic structure of the cue sentence; in this event structure either the agent or the patient is foregrounded, depending on active or passive voice. The conceptual foregrounding of patients is hypothesized to induce a bias in visual attention towards patients in subsequently presented event scenes, leading to an increase in first fixations on patients and a decrease in agent-first fixations. It is important to note that the viewpoint conveyed by actives and passives is independent from lexical semantics and form (e.g., "the man was hugged by the woman" and "the bird was eaten by the cat" converge in their viewpoint). This means that syntactic cueing effects could arise when cue and target event overlap in their most basic conceptual structure (i.e., a core skeleton of "agent acting on patient"), regardless of overlap in event type, and agent/patient identity. In the case of linguistic cues (in this case, entire active and passive sentences) preceding visual stimuli, we expect the syntactic structure of the cue sentences to influence the viewpoint that the perceiver takes on a subsequent unrelated event.

We propose that one can shed light on the process of scene apprehension using a brief exposure paradigm (Dobel et al., 2007; Gerwien & Flecken, 2016; Greene & Oliva, 2009) and eye tracking. In this paradigm, a picture is presented to participants so briefly that they either can only perceive it parafoveally (Dobel et al., 2007; Dobel, Glanemann, Kreysa, Zwitserlood, & Eisenbeiß, 2011) or have time for only a single saccade and fixation on the picture (Gerwien & Flecken, 2016). Target picture presentation times in brief exposure studies range from 37 ms (Hafri et al., 2013) when pictures are presented at the center of the visual field to 300 ms when pictures are presented at the corners of the display and thus require eye movements in order to extract detailed information (Gerwien & Flecken, 2016). Given that programming and executing a saccade takes between 100 and 200 ms (e.g., Kirchner & Thorpe, 2006; Pierce, Clementz, & McDowell, 2019), visual information can only be extracted foveally from the latter kinds of briefly presented stimuli for approximately 100–200 ms.

The location of the first and only fixation in the briefly presented picture is taken to be a direct reflex of the process of event apprehension (Gerwien & Flecken, 2016): Based on parafoveally collected information, viewers identify the core structure of the event and then rapidly decide, e.g., whether to fixate on the agent or the patient in the picture, first. Hence, an analysis of first fixation locations to tap into scene apprehension avoids a reliance on offline measures alone that might be influenced by memory and post-hoc reasoning (Firestone & Scholl, 2016; Lupyan, 2016; Lupyan et al., 2020). We hypothesize that the planning and execution of the first fixation can be influenced by the pre-activated conceptual structure underlying the active or passive cue sentences, including the respective event viewpoint. We hypothesize this reflex of the apprehension process to be the locus of a potential syntactic cueing effect: A linguistically cued event viewpoint should be reflected in what people visually attend to *first* in the event picture.

Here, participants first described a picture of a cue event and then they saw a briefly presented target event. Crucially, cue event descriptions had either an active or a passive sentence structure. After producing the cue sentences, an unrelated target picture appeared for only 300 ms in one of the four screen corners, leaving time for only one fixation on the picture (Fig. 1). Participants then indicated by button press whether a probe picture presented next matched the target picture or not, to ensure participants attended to the target pictures. This design allowed us to test whether entire event representations constructed during speaking can guide the apprehension of subsequently seen events, reflected in cueing effects on the location of the first fixation on target pictures.

2. Methods

2.1. Participants

Forty-one native speakers of Dutch (27 female, age: mean = 24, range = 20–34) from the participant pool of the Max Planck Institute for Psycholinguistics participated for payment. Data from two additional participants were lost due to technical errors in recording or exporting the eye tracking data. The experiment was approved by the ethics committee of the Faculty of Social Sciences at Radboud University Nijmegen.

2.2. Materials and design

Materials consisted of cue, target, and probe pictures. Cue pictures showed 18 different transitive actions with human agents and patients (cf. Appendix A.1, pictures were taken from Segaert, Menenti, Weber, & Hagoort, 2011). Cue pictures were photographed with four actor pairs (two man-woman pairs, two girl-boy pairs) against a black background

[1] Actives and passives also differ on additional dimensions. Passives are less frequent and impose more cognitive load during planning than actives and are morphologically derived, whereas actives are not (Sauppe, 2017a). For the current purpose, however, only the different event viewpoints they entail are relevant.

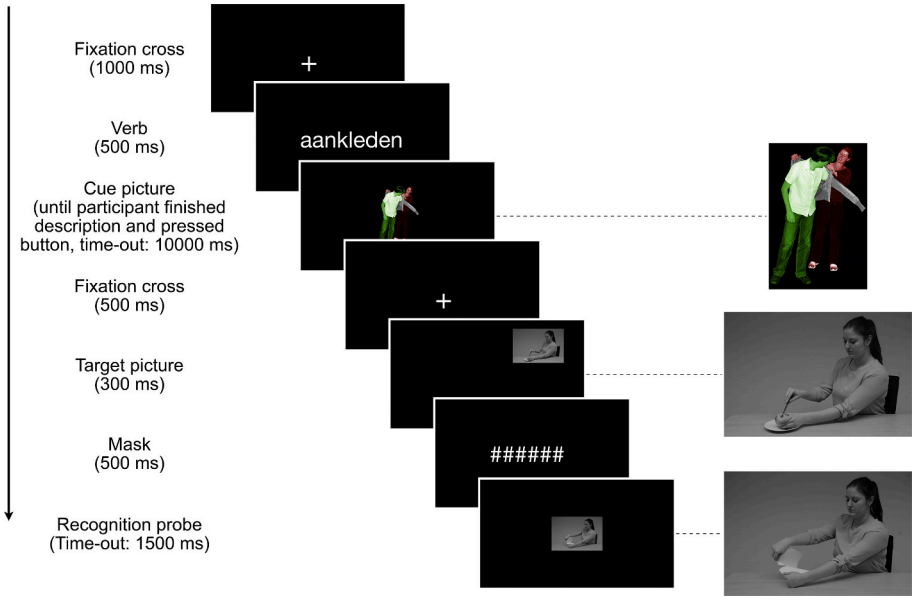

Fig. 1. Trial structure and example stimuli. Trials started with displaying the verb to describe the cue picture (here: "to dress (someone)"). In cue pictures, agent/patient were colored green/red or vice versa; participants were instructed to begin their descriptions with the green character (eliciting active or passive sentences). Cue pictures were presented on the screen until participants pressed a button after having finished their description. Next, after a central fixation cross, target pictures were briefly presented for 300 ms in one of the four screen corners. Finally, a recognition probe was presented and participants indicated by button press whether it matched the target picture. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

(cf. Fig. 1). Agent and patient were colored in red and green and participants were instructed to describe these pictures starting with the green character and using a prespecified verb; this reliably elicited active and passive sentences (as in Segaert et al., 2011; Segaert, Menenti, Weber, Petersson, & Hagoort, 2012).

Target pictures showed 36 transitive events with animate agents and inanimate objects as patients (cf. Appendix A.2, pictures were taken from Sakarias & Flecken, 2019; twenty pictures had female agents). Each target picture appeared eight times over the course of the experiment, once in each of eight blocks (four times after an active and four times after a passive cue event), and in each position on the screen (cf. Fig. 1), with agent-left and agent-right orientation, respectively. Each target picture was paired with eight different cue event pictures, each showing different agent-patient combinations and different actions. One half of the participants saw a given cue-target pair with an active cue, the other half saw it with a passive cue. For each participant, the order of blocks was randomized and the order of trials within blocks was pseudo-randomized, so that no more than two consecutive target pictures appeared in the same screen position.

Probe recognition pictures were taken from the same stimulus pool as target pictures (Sakarias & Flecken, 2019). The probe recognition task had three conditions: target and probe picture were identical (Match condition, half of the trials), the agent mismatched, or the patient/action mismatched (each 25% of the trials). For the Action/Patient Mismatch trials, one of the other target events with the same agent was presented. For the Agent Mismatch trials, pictures of the same event with a different agent were presented.

2.3. Procedure

Participants were tested individually in a laboratory booth. The experiment was programmed in Presentation (Neurobehavioral Systems, Berkeley). Fixation data were collected with a SMI RED250m eye tracker (Sensomotoric Instruments, Teltow), sampling at 250 Hz. Stimuli were displayed on a 15.6″ laptop computer screen with a resolution of 1920 × 1080 pixels, positioned approximately 60 cm away from participants. Target pictures subtended a visual angle of 8.35° horizontally (500 pixels) and 5.64° vertically (333 pixels); the target pictures' center was 9.70° away from the central fixation cross participants fixated on at stimulus onset. Participants first received written instructions on the task and then read further instructions on the screen. After completing six practice trials, they had the chance to ask questions to the experimenter. The eye tracker was then calibrated with a five-point calibration and a four-point validation procedure and participants were told to sit still and not to move their eyes away from the screen. Participants wore a headset recording their descriptions of cue pictures. After every second block there was a self-timed break. The eye tracker was re-calibrated after each break. The total experimental session lasted around 50 min.

2.4. Data processing and analyses

For each target picture, (elliptical) agent and patient areas of interest were defined manually in the eye tracker manufacturer's BeGaze software. The agent area encompassed the face and the upper part (head and part of upper body) of the person performing the action. The patient area encompassed the object being manipulated (i.e., the patient in the narrow sense) and also the agent's hands and a potential instrument (i.e., where the action took place). It is often difficult to separate patients and action regions in naturalistic event depictions, e.g., when the

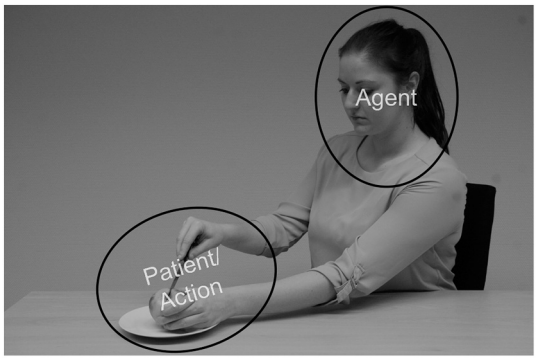

Fig. 2. Example of areas of interest on target stimuli for fixation analyses (areas were not visible to participants).

agent's hands are touching an object. As patients have close ties to

actions (at least in syntax, Kratzer, 1996), we employed an area of interest that encompasses both the patient and the action (Fig. 2).[2] Fixations were detected using the manufacturer's algorithm as implemented in BeGaze.

Trials in which participants did not produce the intended cue sentence (e.g., an active instead of passive when the patient was colored green) or did not look at the target picture during the brief exposure period were excluded from analyses.[3] In addition, two participants who had less than 50% of trials left after exclusions and one participant who had no correct probe recognition trials in the Match condition were excluded. On balance, 9852 trials from 38 participants (84.1% of all data) were available for analyses.

Single-trial level analyses were conducted with *brms* (Bürkner, 2017, 2018; R Core Team, 2018). Fixations to agents and patients/actions during exposure to the target picture were analysed with hierarchical Bayesian Bernoulli regression. The critical predictor was cueing condition (active vs. passive). Nuisance predictors (Sassenhagen & Alday, 2016) were: block in which each trial occurred (reflecting how many passive trials had been encountered), and the orientation (agent left vs. right) and the screen position of target pictures. Agent and patient/action fixations were analysed separately (Barr, 2008). Models included random intercepts and slopes for cue condition by participant and by item, consisted of six chains with 6000 iterations (including 3000 warm-up iterations) and employed Student t distributions (5 degrees of freedom, $\mu = 0$, $\sigma = 3$) as priors for all predictors and the intercept. Predictive model performance with and without the cue condition predictor was assessed using model stacking (Yao, Vehtari, Simpson, & Gelman, 2018). Frequentist hierarchical regressions were computed with *lme4* (Bates, Mächler, Bolker, & Walker, 2015) to supplement the Bayesian analyses and showed the same pattern of results. Statistical significance was assessed with likelihood ratio tests. The maximal random effects structure (that, in the case of frequentist models, allowed convergence) was used for all models (Barr, 2013; Barr, Levy, Scheepers, & Tily, 2013). Categorical predictors were sum coded. Block as continuous predictor was mean-centered.

3. Results

Participants fixated on target pictures on average 200 ms after stimulus onset (SD = 22 ms). Whether the cue pictures were described with actives or passives influenced how participants subsequently viewed pictures during brief exposure (Fig. 3). After passive cues, the likelihood of first fixations on the agent decreased and the likelihood of first fixations on the patient/action of the target events increased, as compared to after active cues. Models including cue condition as a predictor for the likelihood of agent and patient/action fixations performed better in model stacking than models ignoring the cues (Table 1; $p_{Patient/Action} = 0.02$, $\chi^2(1) = 5.38$ and $p_{Agent} = 0.04$, $\chi^2(1) = 4.01$ in frequentist models). In trials in which neither the agent nor patient/action area of interest were fixated, participants mostly fixated the center of the picture in-between these two areas (as in previous studies, e.g., Gerwien & Flecken, 2016). These center fixations were presumably driven by the demands of the recognition task that required participants to rapidly extract information on the entire event. Concerning the two areas of interest, agents were more likely than patients to be fixated first on average, most likely because both agents as such and humans in particular are overall more salient (Cohn & Paczynski, 2013; Crouzet, Kirchner, & Thorpe, 2010; Gao, Baker, Tang, Xu, & Tenenbaum, 2019; Rösler, End, & Gamer, 2017; Webb, Knott, & MacAskill, 2010) and because the human agents in the current stimuli were larger than the inanimate patients.

In the recognition task, responses to the probe were slower and less accurate when either the agent or the patient/action mismatched as compared to when the briefly presented target and the probe pictures matched. Whether the cue sentence had an active or passive structure had no effect on recognition performance (cf. Appendix A.3).

4. Discussion and conclusions

We show that visual event apprehension can be guided by the syntactic structure of recently uttered sentences. Whilst the core event role configuration of cue sentences was kept constant, they differed in the expression of viewpoint on the event—one where either the agent or the patient was foregrounded conceptually. This viewpoint subsequently influenced the attentional prioritization of agents or patients during the planning and execution of the very first fixation onto the briefly presented target event pictures. We take these first fixations to be a direct reflex of the ongoing or possibly finished apprehension process. Participants did likely retrieve the core event structure information parafoveally (Dobel et al., 2007; Hafri et al., 2013), including information on agents and patients and their location (i.e., they extracted what is often called the event's gist, Henderson & Ferreira, 2004). On the basis of this information, they decided where to place their first fixation for further visual information uptake. While the process of event structure extraction itself thus may not have been affected, the subsequent first direction of gaze into the event pictures was informed by the viewpoint conveyed by the syntactic structure of the cue sentences. Event apprehension and saccade programming were executed rapidly: target pictures were fixated already after approximately 200 ms. This means that people could compute their first fixation already after only minimal exposure to the stimuli, and that the cue sentences' syntactic structure thus exerted influence on early perceptual processing stages.

Crucially, cue and target events were unrelated: Whilst cue events involved a human agent and a human patient, target events involved a human agent and an inanimate patient (Fig. 1). The discrepancy in event type and in agent and patient properties (such as animacy), however, still allowed for viewpoint cueing from speaking to seeing. This underlines that the effect took place at the level of the conceptual structure of the events, which includes viewpoint information. The abstract conceptual event structure foregrounding either the agent or the patient was part of the message (Levelt, 1989, 1999)generated during production of the cue sentences (cf. also Bunger, Papafragou, & Trueswell, 2013). We propose that this event structure remained activated also after the sentence was uttered. It could therefore "warp" viewers' event apprehension by exerting a top-down influence on perceivers' decision on which part of event pictures appeared most attention-worthy and should be looked at first under the pressing demands of the task to recognize entire events with only brief exposure (cf. Lupyan, 2012). This process may be similar to the processes underlying syntactic priming during language production (Bock, 1986; Pickering & Ferreira, 2008), where a representation stays active after recent use and influences subsequent processing. The event representation activated during speaking retained activation and was used for subsequent seeing, i.e., when extracting the gist and deciding the starting point for detailed processing of the target event.[4]

The effect of active and passive cue sentences extends conceptual

[2] The minimal distance between areas of interest was on average 1.40° (visual angle: SD = 0.31°, range = 0.78–1.97°; pixels: mean = 82.19 px, SD = 18.09 px, range = 46–116 px). The eye tracker's gaze position accuracy is given as 0.4° by the manufacturer.

[3] There were no trials in which the cue picture display timed out because participants never needed more than 10,000 ms to describe it.

[4] It remains unknown whether abstract event structures activated during comprehension could also influence subsequent apprehension of unrelated events in a similar way, or whether cue and target events would need to be highly similar in order to retain activation of an abstract event structure long enough (cf. lexical boost in syntactic comprehension priming, Branigan, Pickering, & McLean, 2005; Tooley & Traxler, 2010).

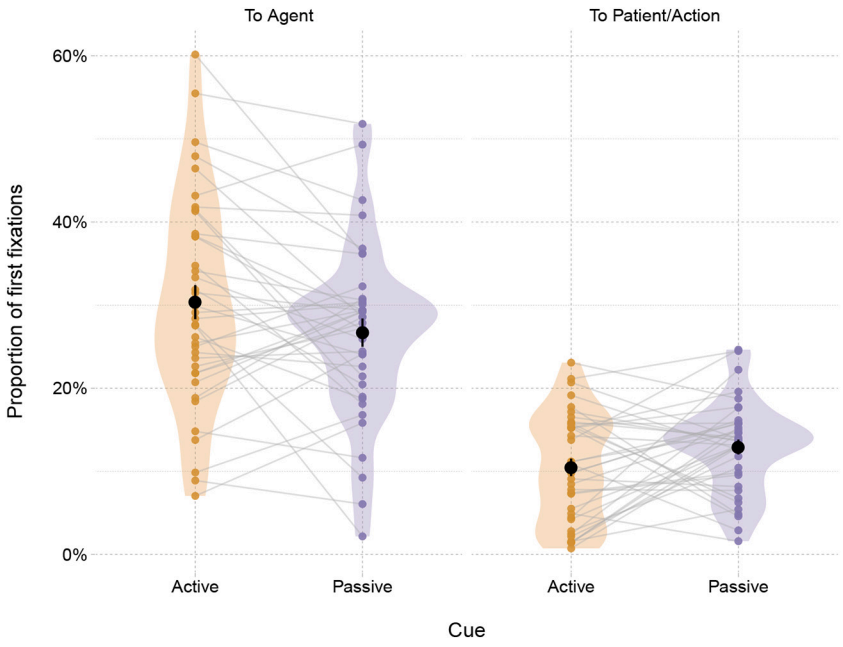

Fig. 3. Proportions of first fixations in agent and patient/action regions in the target-event pictures after active and passive cues. Connected dots and densities represent participant means. Black dots represent means of participant means; error bars indicate one standard error of the mean.

Table 1
Results of hierarchical Bayesian Bernoulli regression predicting the likelihood of fixations on the patient/action and agent regions in briefly exposed target pictures. All Pareto k values < 0.5 (Vehtari, Gelman, & Gabry, 2017).

	Patient/Action Fixations			Agent Fixations		
Parameter	Mean	SD	95% CI	Mean	SD	95% CI
Intercept	−3.89	0.19	[−4.28, −3.52]	−1.47	0.16	[−1.78, −1.15]
Cue (passive)	*0.27*	*0.13*	*[0.03, 0.52]*	*−0.14*	*0.07*	*[−0.28, 0.01]*
Block (centered)	0.01	0.02	[−0.03, 0.05]	0	0.01	[−0.02, 0.03]
Agent position (left)	0.21	0.15	[−0.07, 0.50]	−0.10	0.07	[−0.25, 0.03]
Screen Position (top left)	1.07	0.15	[0.78, 1.36]	−1.22	0.07	[−1.37, −1.08]
Screen Position (top right)	1.14	0.14	[0.86, 1.42]	−1.15	0.07	[−1.29, −1.01]
Screen Position (bottom left)	−1.43	0.24	[−1.95, −1.01]	1.21	0.05	[1.11, 1.32]
Agent position (left) ×Screen Position (top left)	−2.15	−0.15	-[1.86, 2.44]	-1.01	−0.07	-[−1.16, −0.87]
Agent position ×Screen Position (top right)	-2.58	−0.15	-[−2.87, −2.30]	−1.05	−0.07	-[0.92, 1.19]
Agent position ×Screen Position (bottom left)	−1.07	−0.24	-[0.65, 1.59]	-1.11	−0.05	-[−1.21, −1.01]
Model stacking weights:	with Cue		0.951	with Cue		0.786
	without Cue		0.049	without Cue		0.214
Response:	Log odds of fixation to patient/action vs. everywhere else			Log odds of fixation toagent vs. everywhere else		

guidance theories of scene apprehension and eye movements to the domain of events (Henderson, 2017; Henderson et al., 2007; Henderson, Hayes, Peacock, & Rehrig, 2019) and shows how language can provide such conceptual feedback to initial attention allocation. It expands the evidence for language-perception interactions to the realm of sentences and relational categories. To date, it could be shown that labels denoting object concepts facilitate perceptual categorization of these objects. Here, we show that sentences that convey a viewpoint through a syntactic structure can transiently cue the conceptual salience of relational percepts and guide the direction of initial gaze into briefly presented event pictures, resulting in early attentional biases in visual information processing.

Cueing effects of linguistic labels on object perception in the literature were mainly behavioural (with the exception of, e.g., Boutonnet & Lupyan, 2015 and Samaha et al., 2018, who report effects on early visual EEG responses), and assessed post-hoc, e.g., through button presses (cf. Firestone & Scholl, 2016). Here, by contrast, we report an effect on first fixation locations (Gerwien & Flecken, 2016), providing a direct window into event processing and demonstrating that syntactically conveyed event viewpoints play a role in mediating early visual scene processing.

Both active and passive sentences served as appropriate cues for the uptake of information relevant to the task, i.e., extracting agent-patient relations for later recognition,[5] but their differing viewpoints elicited differential *prioritization* in online attention allocation (to either the agent or the patient).

Could the cueing effect, at least in part, be driven by a reliance on verbal encoding of target events (due to the production task or the demands of the recognition task), inducing more early patient fixations for passives (Sauppe, 2017b)? Even though people may rely on verbal strategies to support memory (Trueswell & Papafragou, 2010), such strategies are unlikely to go beyond labelling of event type to include the planning of syntactic alternations, as event viewpoint was irrelevant to the task. Exposure to the pictures for only 300 ms is also likely not sufficient to plan the grammatical structure of entire sentences (Griffin & Bock, 2000), including such syntactic alternations as active and passives. In addition, verbal encoding strategies may not play a role in the encoding of complex scenes for recall (Rehrig, Hayes, Henderson, & Ferreira, 2020).

[5] Note that in the current study both agents and patients were always overtly mentioned in cue sentences and only differed in foregrounding and backgrounding through syntactic structure.

In conclusion, cueing effects of grammatical structure on event processing open up a new range of possibilities for exploring language-perception interactions, beyond features of single words (like gender, Sato & Athanasopoulos, 2018), and making use of linguistic diversity (Norcliffe, Harris, & Jaeger, 2015). Visual event apprehension could, for example, also be modulated by other grammatical phenomena that attune agent and patient salience such as differential subject and object marking (de Hoop & Malchukov, 2008), ergativity (Bickel, Witzlack-Makarevich, Choudhary, Schlesewsky, & Bornkessel-Schlesewsky, 2015; Dixon, 1994), or information-structurally driven word order variations (Downing & Noonan, 1995).

Author contributions

Both authors contributed equally to the work reported here.

Data availability

Raw data and analysis scripts are available from https://osf.io/3gvch/.

Acknowledgements

This work was funded by Swiss National Science Foundation grants 100015_160011 and 100015_182845 and the Department of Comparative Language Science, University of Zurich (S.S.), and the Neurobiology of Language Department, Max Planck Institute for Psycholinguistics (M. F.). The authors thank Balthasar Bickel, Peter Hagoort, Arrate Isasi-Isasmendi, Ksenija Slivac, Julia Misersky, as well as three reviewers for helpful comments on a previous version of the manuscript. The authors further thank Katrien Segaert for providing the stimuli for the cue pictures, Anna Jancso for programming help, and Giuachin Kreiliger for statistical consultation.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2020.104516.

References

Barr, D. J. (2008). Analyzing "visual world" eyetracking data using multilevel logistic regression. *Journal of Memory and Language, 59*(4), 457–474. https://doi.org/10.1016/j.jml.2007.09.002.

Barr, D. J. (2013). Random effects structure for testing interactions in linear mixed-effects models. *Frontiers in Psychology, 4*(328). https://doi.org/10.3389/fpsyg.2013.00328.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. Retrieved from http://CRAN.R-project.org/package=lme4 10.18637/jss.v067.i01.

Bickel, B., Witzlack-Makarevich, A., Choudhary, K. K., Schlesewsky, M., & Bornkessel-Schlesewsky, I. (2015). The neurophysiology of language processing shapes the evolution of grammar: Evidence from case marking. *PLoS One, 10*(8), Article e0132819. https://doi.org/10.1371/journal.pone.0132819.

Biederman, I. (1981). On the semantics of a glance at a scene. In M. Kubovy, & J. R. Pomerantz (Eds.), *Perceptual organization* (pp. 213–253). London: Routledge. https://doi.org/10.4324/9781315512372.

Bock, K. (1986). Syntactic persistence in language production. *Cognitive Psychology, 18*(3), 355–387. https://doi.org/10.1016/0010-0285(86)90004-6.

Bock, K., Irwin, D. E., & Davidson, D. J. (2004). Putting first things first. In J. M. Henderson, & F. Ferreira (Eds.), *The interface of language, vision, and action: Eye movements and the visual world* (pp. 249–278). New York/Hove: Psychology Press.

Boutonnet, B., & Lupyan, G. (2015). Words jump-start vision: A label advantage in object recognition. *Journal of Neuroscience, 35*(25), 9329–9335. https://doi.org/10.1523/JNEUROSCI.5111-14.2015.

Branigan, H. P., Pickering, M. J., & McLean, J. F. (2005). Priming prepositional-phrase attachment during comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(3), 468–481. https://doi.org/10.1037/0278-7393.31.3.468.

Bunger, A., Papafragou, A., & Trueswell, J. C. (2013). Event structure influences language production: Evidence from structural priming in motion event description. *Journal of Memory and Language, 69*(3), 299–323. https://doi.org/10.1016/j.jml.2013.04.002.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software, 80*(1). https://doi.org/10.18637/jss.v080.i01.

Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *R Journal, 10*(1), 395–411. https://doi.org/10.32614/RJ-2018-017.

Castelhano, M. S., & Henderson, J. M. (2007). Initial scene representations facilitate eye movement guidance in visual search. *Journal of Experimental Psychology: Human Perception and Performance, 33*(4), 753–763. https://doi.org/10.1037/0096-1523.33.4.753.

Cohn, N., & Paczynski, M. (2013). Prediction, events, and the advantage of agents: The processing of semantic roles in visual narrative. *Cognitive Psychology, 67*(3), 73–97. https://doi.org/10.1016/j.cogpsych.2013.07.002.

Core Team, R. (2018). R: A language and environment for statistical computing [Computer software manual]. Vienna. Retrieved from http://www.R-project.org.

Crouzet, S. M., Kirchner, H., & Thorpe, S. J. (2010). Fast saccades toward faces: Face detection in just 100 ms. *Journal of Vision, 10*(4), 16. https://doi.org/10.1167/10.4.16.

de Hoop, H., & Malchukov, A. L. (2008). Case-marking strategies. *Linguistic Inquiry, 39*(4), 565–587. https://doi.org/10.1162/ling.2008.39.4.565.

Dixon, R. M. W. (1994). *Ergativity*. Cambridge: Cambridge University Press.

Dobel, C., Gumnior, H., Bölte, J., & Zwitserlood, P. (2007). Describing scenes hardly seen. *Acta Psychologica, 125*(2), 129–143. https://doi.org/10.1016/j.actpsy.2006.07.004.

Dobel, C., Glanemann, R., Kreysa, H., Zwitserlood, P., & Eisenbeiß, S. (2011). Visual encoding of coherent and non-coherent scenes. In J. Bohnemeyer, & E. Pederson (Eds.), *Vol. 11. Event representation in language and cognition* (pp. 189–215). Cambridge: Cambridge University Press.

Downing, P. A., & Noonan, M. (Eds.). (1995). *Vol. 30. Word order in discourse*. Amsterdam/Philadelphia: John Benjamins.

Dowty, D. (1991). Thematic proto-roles and argument selection. *Language, 67*(3), 547–619. https://doi.org/10.1353/lan.1991.0021.

Firestone, C., & Scholl, B. J. (2016). Cognition does not affect perception: Evaluating the evidence for "top-down" effects. *Behavioral and Brain Sciences, 39*, Article e229. https://doi.org/10.1017/S0140525X15000965.

Galazka, M., & Nyström, P. (2016). Infants' preference for individual agents within chasing interactions. *Journal of Experimental Child Psychology, 147*, 53–70. https://doi.org/10.1016/j.jecp.2016.02.010.

Gao, T., Baker, C. L., Tang, N., Xu, H., & Tenenbaum, J. B. (2019). The cognitive architecture of perceived animacy: Intention, attention, and memory. *Cognitive Science: A Multidisciplinary Journal, 43*(8), Article e12775. https://doi.org/10.1111/cogs.12775.

Gerwien, J., & Flecken, M. (2016). First things first? Top-down influences on event apprehension. In A. Papafragou, D. Grodner, D. Mirman, & J. C. Trueswell (Eds.), *Proceedings of the 38th annual meeting of the Cognitive Science Society* (pp. 2633–2638). Cognitive Science Society (Austin, TX).

Gilbert, C. D., & Li, W. (2013). Top-down influences on visual processing. *Nature Reviews Neuroscience, 14*, 350–363. https://doi.org/10.1038/nrn3476.

Glanemann, R., Zwitserlood, P., Bölte, J., & Dobel, C. (2016). Rapid apprehension of the coherence of action scenes. *Psychonomic Bulletin & Review, 23*, 1566–1575. https://doi.org/10.3758/s13423-016-1004-y.

Greene, M. R., & Oliva, A. (2009). The briefest of glances: The time course of natural scene understanding. *Psychological Science, 20*(4), 464–472. https://doi.org/10.1111/j.1467-9280.2009.02316.x.

Griffin, Z. M., & Bock, K. (2000). What the eyes say about speaking. *Psychological Science, 11*(4), 274–279. https://doi.org/10.1111/1467-9280.00255.

Hafri, A., Papafragou, A., & Trueswell, J. C. (2013). Getting the gist of events: Recognition of two-participant actions from brief displays. *Journal of Experimental Psychology: General, 142*(3), 880–905. https://doi.org/10.1037/a0030045.

Hafri, A., Trueswell, J. C., & Strickland, B. (2018). Encoding of event roles from visual scenes is rapid, spontaneous, and interacts with higher-level visual processing. *Cognition, 175*, 36–52. https://doi.org/10.1016/j.cognition.2018.02.011.

Henderson, J. M. (2017). Gaze control as prediction. *Trends in Cognitive Sciences, 21*(1), 15–23. https://doi.org/10.1016/j.tics.2016.11.003.

Henderson, J. M., & Ferreira, F. (2004). Scene perception for psycholinguists. In J. M. Henderson, & F. Ferreira (Eds.), *The interface of language, vision, and action: Eye movements and the visual world* (pp. 1–58). New York/Hove: Psychology Press.

Henderson, J. M., Brockmole, J. R., Castelhano, M. S., & Mack, M. (2007). Visual saliency does not account for eye movements during visual search in real-world scenes. In R. P. G. van Gompel, M. H. Fischer, W. S. Murray, & R. L. Hill (Eds.), *Eye movements: A window on mind and brain* (pp. 537–562). Oxford: Elsevier.

Henderson, J. M., Hayes, T. R., Peacock, C. E., & Rehrig, G. (2019). Meaning and attentional guidance in scenes: A review of the meaning map approach. *Vision, 3*(2), 19. https://doi.org/10.3390/vision3020019.

Hollingworth, A., & Henderson, J. M. (1998). Does consistent scene context facilitate object perception? *Journal of Experimental Psychology: General, 127*(4), 398–415. https://doi.org/10.1037/0096-3445.127.4.398.

Jackendoff, R. (1990). *Semantic structures*. Cambridge, MA/London: MIT Press.

Johnson, S. C. (2003). Detecting agents. *Philosophical Transactions of The Royal Society B: Biological Sciences, 358*(1431). https://doi.org/10.1098/rstb.2002.1237.

Kazenin, K. I. (2001). The passive voice. In M. Haspelmath, E. König, W. Oesterreicher, & W. Raible (Eds.), *vol. 20/2. Language typology and language universals: An international handbook* (pp. 899–916). Berlin/New York: Walter de Gruyter. https://doi.org/10.1515/9783110171549.2.

Keenan, E. L., & Dryer, M. S. (2007). Passive in the world's languages. In T. Shopen (Ed.), *Language typology and syntactic description. Volume I: Clause structure* (2nd ed., pp. 325–361). Cambridge: Cambridge University Press.

Kirchner, H., & Thorpe, S. J. (2006). Ultra-rapid object detection with saccadic eye movements: Visual processing speed revisited. *Vision Research, 46*(11), 1762–1776. https://doi.org/10.1016/j.visres.2005.10.002.

Kratzer, A. (1996). Severing the external argument from its verb. In J. Rooryck, & L. Zaring (Eds.), *Vol. 33. Phrase structure and the lexicon* (pp. 109–137). Dordrecht: Springer. https://doi.org/10.1007/978-94-015-8617-7.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.

Levelt, W. J. M. (1999). Producing spoken language: A blueprint of the speaker. In C. M. Brown, & P. Hagoort (Eds.), *The Neurocognition of language* (pp. 83–122). Oxford: Oxford University Press.

Lupyan, G. (2012). Linguistically modulated perception and cognition: The label-feedback hypothesis. *Frontiers in Psychology, 3*(54). https://doi.org/10.3389/fpsyg.2012.00054.

Lupyan, G. (2016). Not even wrong: The "it's just x" fallacy. *Behavioral and Brain Sciences, 39*. https://doi.org/10.1017/S0140525X15002721. e251.

Lupyan, G., & Ward, E. J. (2013). Language can boost otherwise unseen objects into visual awareness. *Proceedings of the National Academy of Sciences of the United States of America, 110*(35), 14196–14201. https://doi.org/10.1073/pnas.1303312110.

Lupyan, G., Abdel Rahman, R., Boroditsky, L., & Clark, A. (2020). Effects of language on visual perception. *Trends in Cognitive Sciences, 24*(11), 930–944. https://doi.org/10.1016/j.tics.2020.08.005.

Norcliffe, E., Harris, A. C., & Jaeger, T. F. (2015). Cross-linguistic psycholinguistics and its critical role in theory development: Early beginnings and recent advances. *Language, Cognition and Neuroscience, 30*(9), 1009–1032. https://doi.org/10.1080/23273798.2015.1080373.

Ostarek, M., & Huettig, F. (2017). Spoken words can make the invisible visible — Testing the involvement of low-level visual representations in spoken word processing. *Journal of Experimental Psychology: Human Perception and Performance, 43*(3), 499–508. https://doi.org/10.1037/xhp0000313.

Pickering, M. J., & Ferreira, V. S. (2008). Structural priming: A critical review. *Psychological Bulletin, 134*(3), 427–459. https://doi.org/10.1037/0033-2909.134.3.427.

Pierce, J. E., Clementz, B. A., & McDowell, J. E. (2019). Saccades: Fundamentals and neural mechanisms. In C. Klein, & U. Ettinger (Eds.), *Eye movement research. An introduction to its scientific foundations and applications* (pp. 11–71). Cham: Springer. https://doi.org/10.1007/978-3-030-20085-5_2.

Potter, M. C., & Levy, E. I. (1969). Recognition memory for a rapid sequence of pictures. *Journal of Experimental Psychology: General, 81*(1), 10–15. https://doi.org/10.1037/h0027470.

Rehrig, G., Hayes, T. R., Henderson, J. M., & Ferreira, F. (2020). When scenes speak louder than words: Verbal encoding does not mediate the relationship between scene meaning and visual attention. *Memory & Cognition*. https://doi.org/10.3758/s13421-020-01050-4.

Richmond, L. L., & Zacks, J. M. (2017). Constructing experience: Event models from perception to action. *Trends in Cognitive Sciences, 21*(12), 962–980. https://doi.org/10.1016/j.tics.2017.08.005.

Rissman, L., & Majid, A. (2019). Thematic roles: Core knowledge or linguistic construct? *Psychonomic Bulletin & Review, 26*, 1850–1869. https://doi.org/10.3758/s13423-019-01634-5.

Rissman, L., Woodward, A. L., & Goldin-Meadow, S. (2019). Occluding the face diminishes the conceptual accessibility of an animate agent. *Language, Cognition and Neuroscience, 34*(3), 273–288. https://doi.org/10.1080/23273798.2018.1525495.

Rösler, L., End, A., & Gamer, M. (2017). Orienting towards social features in naturalistic scenes is reflexive. *PLoS One, 12*(7), Article e0182037. https://doi.org/10.1371/journal.pone.0182037.

Sakarias, M., & Flecken, M. (2019). Keeping the result in sight and mind: General cognitive principles and language-specific influences in the perception and memory of resultative events. *Cognitive Science, 43*(1), Article e12708. https://doi.org/10.1111/cogs.12708.

Samaha, J., Boutonnet, B., Postle, B. R., & Lupyan, G. (2018). Effects of meaningfulness on perception: Alpha-band oscillations carry perceptual expectations and influence early visual responses. *Scientific Reports, 8*, 6606. https://doi.org/10.1038/s41598-018-25093-5.

Sassenhagen, J., & Alday, P. M. (2016). A common misapplication of statistical inference: Nuisance control with null-hypothesis significance tests. *Brain and Language, 162*, 42–45. https://doi.org/10.1016/j.bandl.2016.08.001.

Sato, S., & Athanasopoulos, P. (2018). Grammatical gender affects gender perception: Evidence for the structural-feedback hypothesis. *Cognition, 176*, 220–231. https://doi.org/10.1016/j.cognition.2018.03.014.

Sauppe, S. (2017a). Symmetrical and asymmetrical voice systems and processing load: Pupillometric evidence from sentence production in Tagalog and German. *Language, 93*(2), 288–313. https://doi.org/10.1353/lan.2017.0015.

Sauppe, S. (2017b). Word order and voice influence the timing of verb planning in German sentence production. *Frontiers in Psychology, 8*(1648). https://doi.org/10.3389/fpsyg.2017.01648.

Schyns, P. G., & Oliva, A. (1994). From blobs to boundary edges: Evidence for time- and spatial-scale-dependent scene recognition. *Psychological Science, 5*(4), 195–200. https://doi.org/10.1111/j.1467-9280.1994.tb00500.x.

Segaert, K., Menenti, L., Weber, K., & Hagoort, P. (2011). A paradox of syntactic priming: Why response tendencies show priming for passives, and response latencies show priming for actives. *PLoS One, 6*(10), Article e24209. https://doi.org/10.1371/journal.pone.0024209.

Segaert, K., Menenti, L., Weber, K., Petersson, K. M., & Hagoort, P. (2012). Shared syntax in language production and language comprehension — An fMRI study. *Cerebral Cortex, 22*(7), 1662–1670. https://doi.org/10.1093/cercor/bhr249.

Spelke, E. S., & Kinzler, K. D. (2007). Core knowledge. *Developmental Science, 10*(1), 89–96. https://doi.org/10.1111/j.1467-7687.2007.00569.x.

Summerfield, C., & de Lange, F. P. (2014). Expectation in perceptual decision making: Neural and computational mechanisms. *Nature Reviews Neuroscience, 15*, 745–756. https://doi.org/10.1038/nrn3838.

Summerfield, C., Egner, T., Greene, M., Koechlin, E., Mangels, J., & Hirsch, J. (2006). Predictive codes for forthcoming perception in the frontal cortex. *Science, 314*(5803), 1311–1314. https://doi.org/10.1126/science.1132028.

Tooley, K. M., & Traxler, M. J. (2010). Syntactic priming effects in comprehension: A critical review. *Lang & Ling Compass, 4*(10), 925–937. https://doi.org/10.1111/j.1749-818X.2010.00249.x.

Trueswell, J. C., & Papafragou, A. (2010). Perceiving and remembering events cross-linguistically: Evidence from dual-task paradigms. *Journal of Memory and Language, 63*(1), 64–82. https://doi.org/10.1016/j.jml.2010.02.006.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing, 27*, 1413–1432. https://doi.org/10.1007/s11222-016-9696-4.

Webb, A., Knott, A., & MacAskill, M. R. (2010). Eye movements during transitive action observation have sequential structure. *Acta Psychologica, 133*(1), 51–56. https://doi.org/10.1016/j.actpsy.2009.09.001.

Yao, Y., Vehtari, A., Simpson, D., & Gelman, A. (2018). Using stacking to average Bayesian predictive distributions (with discussion). *Bayesian Analysis, 13*(3), 917–1007. https://doi.org/10.1214/17-BA1091.

Zacks, J. M. (2020). Event perception and memory. *Annual Review of Psychology, 71*, 165–191. https://doi.org/10.1146/annurev-psych-010419-051101.