Searching for Events:
Rapid visual extraction of language-compatible event representations

Junyi Chen (aprchen@sas.upenn.edu)
Department of Psychology, University of Pennsylvania,
425 S University Ave, Philadelphia, PA 19139 USA

John Trueswell (trueswel@psych.upenn.edu)
Department of Psychology, University of Pennsylvania,
425 S University Ave, Philadelphia, PA 19139 USA

Abstract

How does the visual system recognize human interactions, and what is the nature of these representations? Past work suggests observers can automatically recognize event category (e.g., kicking) and event role information (Agent, Patient) from brief displays, so-called rapid gist extraction of event structure. Questions remain though about how quickly event representations are computed and when they might interface with linguistic/cognitive systems. We explored these issues using linguistically guided visual search. Participants' eye movements were recorded as they heard spoken input (e.g., "The red person is kicking the blue person") and searched for the matching image. By manipulating visual preview time prior to hearing the critical verb, we can estimate an upper bound of when visually recognized event information is available to ongoing linguistic processes. And by manipulating the posture of the humans in these images, we can help clarify how event representations are recognized and refined over time.

Keywords: Event cognition, Event roles, Eye-tracking, Real time language processing

Introduction

How are human actions and interactions perceived and categorized as events, composed of event participants? One possibility is that visual processing offers only lower-level information about the properties and locations of entities in the scene; further visual interrogation and active engagement of inference would be needed to 'stitch together' relational information, perhaps via some set of active 'visual routines' (Ullman, 1996). Yet, the primacy of understanding the behavior of conspecifics motivates the idea that the recognition of event structure may be a rapid and spontaneous component of visual perception.

Some initial evidence in support of this view comes from Hafri et al. (2013), who found that participants were above chance at confirming event category (kicking, pushing) and event role information (kicker, kickee) from masked visual images displayed for less than 100 ms (for related findings, see Dobel et al., 2007; Glanemann et al., 2016). This suggests that events and event structure can be computed from a single fixation – i.e., rapid gist extraction of events. Consistent with work in computer vision on event perception (e.g., Sun et al., 2022), participants' identification of events and event roles in Hafri et al. was supported by features of physical posture of the humans in the images (e.g., facing another individual, active posture, outstretched appendages support Agency; see also Vettori et al., 2024; De Freitas & Hafri, 2024). Strikingly, these recognition processes appear to occur spontaneously, even when observers are engaged in a task that does not require the recognition of the events (e.g., Hafri et al., 2018; Ji & Scholl, 2024). For example, when identifying on which side of the screen a red-shirted person appeared across multiple trials, participants were slowed when the event role of this individual changed across trials (Agent to Patient) as compared to when it remained the same (the role switch cost; Hafri et al., 2018), suggesting that event role information was spontaneously encoded.

These findings raise a broader issue about the nature of event representations and how they may be designed to interface with higher level cognitive systems, including language. Notably, the role switch cost observed by Hafri et al. occurred even when the events were different across trials (e.g., kicking then lifting). This suggests that *event independent* representations exist for event roles (Agent, Patient). This is a striking claim, given the primacy of the notion of event roles in linguistics, reasoning, and cognitive development. It is possible, then, that automatic visual processes provide a sketch of the event structure in a format amenable to higher level cognition and language.

Despite these advances, a great deal is yet to be understood. In particular, how quickly is visually- recognized event information available to ongoing linguistic processes? Visual-World eyetracking research suggests that the recognition of a verb can lead to eye fixations on event-relevant participants within 500 ms of the onset of the verb (e.g., Altman & Kamide, 1999). Considering the approximate 100 ms lag for eye movements to respond to linguistic input in the visual world paradigm (Altmann, 2011), it would suggest language interfaces with these representations very rapidly (400 ms or less). However, almost all visual world eye tracking studies provide image preview time to interrogate the scene, permitting time to compute abstract representations of humans, objects, and their affordances (see Huettig et al., 2011, for discussion). Thus, Visual-World eyetracking results suggest that the linguistic processing of verb information is fast, but it does not speak to the time-course of computing event structure from visual input.

One hint regarding the speed of the visual recognition of event structure comes from the Hafri et al. (2018) study of role switch costs. As noted by the authors, the average response time to identify the side of the red-shirted person

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

was approximately 400 ms. Given that role switch cost effects were observed from these data, it sets the upper bound on the time required to compute such information. However, this finding does not speak to when this information is available to linguistic processing. It may also be that role switch cost effects are due to conflicting intermediate representations that serve to later compute event structure, raising the possibility that they are not completed computations accessible to ongoing linguistic processes.

We explore these issues in studies of linguistically-guided visual search. In two experiments, participants' eye movements were recorded as they heard spoken input (e.g., "The red person is kicking the blue person") and searched for the matching image. By manipulating visual preview time prior to hearing the critical verb in the sentence, we can estimate an upper bound of when visually recognized event information is available to ongoing linguistic processes. And by manipulating the physical posture of the humans in the target and foil images in ways relevant to the recognition of events and event roles, we can help clarify how event representations are recognized and refined over time.

Based on the research reviewed above, we predict that: (1) if event gist extraction makes rapid contact with language processing, participants should show above-chance target identification upon hearing the verb even without visual preview, and (2) the typicality of postures in the Target Action (Experiment 1) and Foil Action (Experiments 1 and 2) should modulate target identification, with typical Patient postures facilitating event recognition.

Experiment 1

In Exps 1A and 1B, participants heard a spoken description (e.g., "The red person is kicking the blue person") while viewing two images. One action matched the sentence (Target), the other did not (Foil), and participants were to locate the Target by clicking on it.

In Exp 1A, the images appeared at the onset of hearing the spoken verb (e.g., "kicking") and we assessed when target-looking was above chance. We expect evidence of rapid identification of events even without image preview. Based on the work reviewed above, rapid visual extraction of event information occurs within 400 ms of Image Onset and therefore could be available to ongoing linguistic processing at that time (i.e., 400 ms post Verb Onset).

In Exp 1B, the images instead appeared at Sentence Onset, providing a preview of the actions before hearing the verb. Preview should permit early encoding of event information and accelerate looks to the Target at the moment of hearing the spoken verb. Note that hearing the "The red person is…" during the preview stage should not help observers predict which image is the Target because on all trials the Agent in both image pairs was the same person (e.g., both were the red shirted person). Thus, image preview allows for the visual encoding of the actions, but linguistic input does not permit prediction of which is the Target. One has to wait to hear the verb (e.g., "kicking").

In both Exp 1A and 1B, we manipulated the postural properties of the Target image (Fig1) to examine their effects on search. For half the trials, the Target had a Patient in a typical Patient-like stance (facing away, arms down). On the other half, the Target had a Patient in a typical Agent-like stance (facing toward, arms out). Target searching should be facilitated when the Target has a Patient in a Patient-like stance, as it will facilitate recognition of the action (kicking). The Foil images were manipulated using a fully crossed design so that, for half of the images, the Foil depicted a Patient in a Patient-like stance, and for the other half, the Foil depicted the Patient in an Agent-like stance. This could have a similar effect on search: a Patient-like Patient will facilitate recognition of the Foil action (tapping) and facilitate the exclusion of it in favor of the Target.

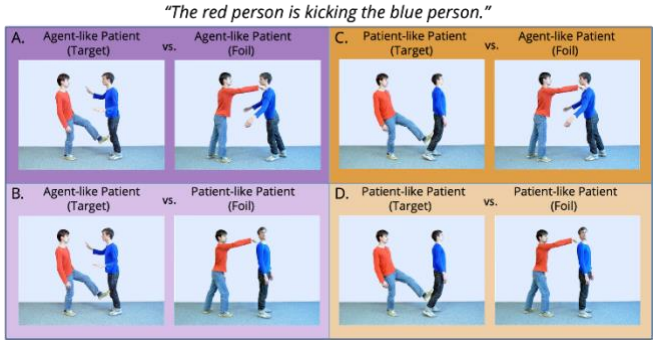

Figure 1: Example item, four conditions (Exp 1). Foil images always had a Different Action than the Target. Target images had an Agent-like Patient (A and B) or a Patient-like Patient (C and D). Foil images had an Agent-like Patient (A and C) or a Patient-like Patient (B and D). Color backgrounds were not shown to participants but correspond to graphed conditions in Fig 2.

Methods

Participants Ninety-one native English speakers at the University of Pennsylvania participated for course credit. Forty participants from Exp 1A and forty from 1B were included in the analyses. Eleven participants were excluded due to being non-native speakers of English (4), high track loss (3), low clicking accuracy (2), repetitive list (1), and failing to calibrate (1).

Materials Stimuli consisted of paired images from Hafri et al. (2013), with each image depicting a two-participant interaction involving persons wearing red and blue shirts (see Fig1 for examples). Eight experimental lists were constructed to counterbalance multiple factors across test conditions, including the side of the Target image (left/right) and the color of the Target Agent (red/blue). Color of the Agent in the Foil image was always the same as the Agent color of the Target. Additionally, the side of the red and the blue person remained consistent within each trial. Each trial included an

audio description structured as *'The red/blue person is verb-ing the blue/red person.*'

Each list consisted of 32 trials distributed across two blocks, with 16 distinct actions presented in Block 1 and repeated (but in a different condition) in a randomized order in Block 2. On each trial, two images were presented at Verb Onset - a Target and a Foil image. An example item depicting the four experimental conditions appears in Fig1.

Procedure Participants were randomly assigned to one of eight stimuli lists. Eye movements were recorded using a Tobii TX300 Pro eye-tracker at a 60 Hz sampling rate. The experiment began with eye-tracker calibration for each participant. Each trial followed the same sequence: a fixation cross appeared in the center of the screen, and participants clicked on the cross to initiate the trial. Upon clicking, participants heard an auditory description and clicked on the matching picture from the display. Participants were instructed to make selections as quickly as possible while remaining accurate. In Exp 1A, images appeared at Verb Onset; in Exp 1B, they appeared at Sentence Onset.

Analysis Gaze location was coded based on screen-side (Target vs. Foil), with the screen divided into equal left and right regions. For each time point, we coded whether participants were looking at the Target side (1) or not (0). For analyses, we created time sequences from 1000 ms before to 2000 ms after audio onset, with data binned into 100 ms intervals. We generated a time course plot examining proportions of looks to Target under all condition combinations (see Fig2).

To examine the statistical reliability of these patterns, we conducted cluster-based permutation analyses using linear mixed-effects models. Looking behavior was quantified using empirical logit (Elogit) transformation of the proportion of looks to Target. Our model tested the effects of Target Patient Posture (Patient- vs. Agent-like), Foil Patient Posture (Patient- vs. Agent-like), and Block (1 vs. 2), including their interactions. The model included random intercepts for subjects and items. For each fixed effect and interaction, neighboring t-tests exceeding 1.5 were summed and designated as clusters. Statistical significance of clusters was assessed using 1000 permutations of conditions, using jlmerclusterperm in R.

Results Exp 1A (Images at Verb Onset)

In Exp 1A, participants were not shown image pairs until the onset of the spoken verb and thus had no preview of images. Eye gaze was analyzed for correct trials only (98% of trials). Fig2 (top panel) plots the proportion of looks to the Target side of the screen as a function of time, relative to the onset of the verb in the spoken sentence. Descriptively, participants show increased Target looks at about 400-500 ms. Although some differences exist across conditions, they appear to be relatively small.

Time to Access Event Information To begin our analyses, we simply ask when participants show reliable above chance looking to the Target side. We interpret this measure as an estimate of when visually-recognized event information is accessible to ongoing linguistic processing. Five mixed-effects linear models were applied to the first five 100 ms samples of the eye movement record from the onset of the verb of the form ElogitTargetLooks ~ 1 + (1|Subject) + (1|Item). A reliably positive intercept would be evidence of above chance looking to the Target in that sample because the Elogit of a 0.5 probability is 0.0. The intercept became reliably positive (above chance) during the 400-500 ms sample (Est. = 0.266, t = 2.85, p = 0.012). Assuming a minimum of a 100 ms lag for eye movements to respond to linguistic input (Altmann, 2011), this means that event recognition processes interface with linguistic processes approximately 300-400 ms into image (and verb) processing, which is during the auditory perception of the verb (average verb duration of our stimuli is 463 ms).

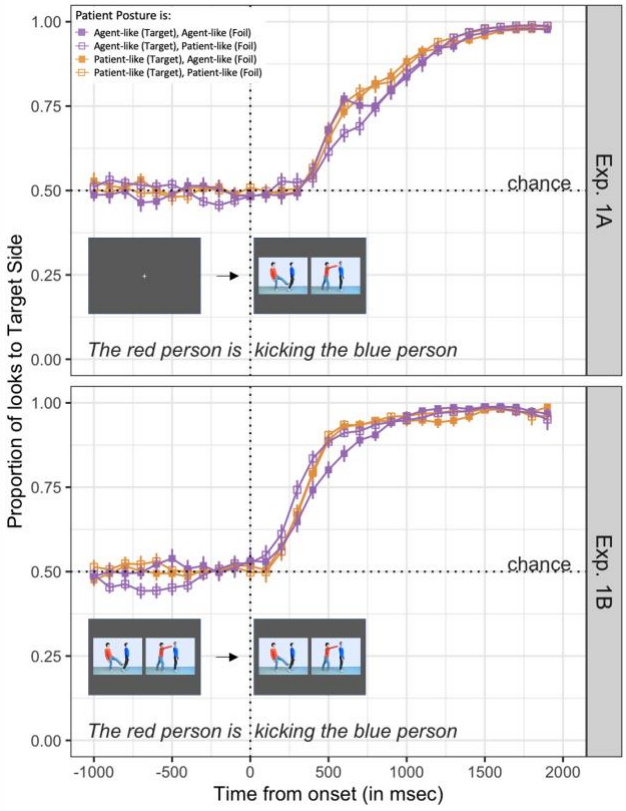

Figure 2: Proportion of looks to Target Side as a function of Time from Verb Onset. Means and SE of subject means. Images illustrate the timing of onsets.

Cluster-based Permutation Tests The results of a cluster-based permutation analysis were designed to identify reliable effects of Target Patient Posture (Patient- vs. Agent-like), Foil Patient Posture (Patient- vs. Agent-like), and Block (1 vs. 2), and all possible interactions. Results of the analysis (Table 1A) revealed a marginally significant effect of the Target Patient Posture from 700-1100 ms (p = .077),

reflecting greater looks to the Target if the Patient had Patient-like (i.e., congruent) features. In addition, a reliable main effect of Block was observed from 500-1700 ms (p = .001), reflecting improved Target searching in Block 2. No other clusters of main effects or interactions were significant (all p's > .1).

Table 1: Exp 1 cluster-based permutation tests results.

A.Effects (Exp1A)	Cluster in ms	Sum t	p =
Target Patient Posture	700 to 1100	-11.70	0.077
Block (1 vs. 2)	500 to 1700	-42.99	0.001
B.Effects (Exp1B)	Time of Cluster	Sum t	p =
Target Patient Posture	500 to 800	-10.93	0.071
Block (1 vs. 2)	300 to 800	-16.93	0.001
	1100 to 1400	-14.30	0.001
C.Effects (Combined Exp1A and Exp1B)	Time of Cluster	Sum t	p =
Experiment (1A vs. 1B)	-100 to 1300	90.69	0.001
Target Patient Posture	500 to 900	-14.61	0.029
Foil Patient Posture x Experiment	400 to 700	-8.80	0.001
Target Patient Post. x Foil Patient Post. x Exp.	-700 to -500	5.97	0.001
	400 to 700	-10.01	0.001
Block (1 vs. 2)	400 to 1600	-53.59	0.001
Exp. x Target Patient Post. x Block	1300 to 1500	5.01	0.001
Foil Patient Post. x Block	-600 to -500	-3.81	0.001

(No other reliable effects or interactions)

Formula (Exp 1A&1B): Elog ~ 1 + TargPatPost*FoilPatPost*Block + (1 | Subject) + (1 | TargetVerb)
Formula (Combined): Elog ~ 1 + Exp*TargPatPost*FoilPatPost*Block + (1 | Subject) + (1 | TargetVerb)
Applied to 100 bins from -1000 to 2000. Significant clusters identified with R package jlmerclusterperm

Results Exp 1B (Images at Sentence Onset)

In Exp 1B participants were shown image pairs at Sentence Onset, providing image preview prior to hearing the verb. Eye gaze was analyzed for correct trials only (94% of trials). Fig2 (bottom panel) plots the proportion of looks to the Target relative to the onset of the verb in the spoken sentence. Descriptively, participants show increased Target looks at about 200-300 ms. Although some differences exist across conditions, they appear to be relatively small.

Time to Access Event Information Above chance Target looking was assessed in the same way as Exp 1A. With visual preview, participants showed reliably above chance looks to the Target image very early, starting during the 200-300 ms sample (Est. = 0.423, t = 5.56, p < 0.001). Thus, allowing for an eye movement lag, 100-200 ms of acoustic input was needed to begin to access the event information. Note that because images are onscreen prior to the verb, some subjects were already fixating the Target at Verb Onset. Thus, the striking speed in response to visual input likely reflects cancellation of an eye movement in response to spoken input, which would be faster than planning an eye movement.

Cluster-based Permutation Analysis Results of the cluster-based permutation analysis (Table 1B) revealed a marginally significant effect of the Target Patient Posture from 500-800 ms (p = .071), reflecting greater Target looks if the Patient had Patient-like (i.e., congruent) features. In addition, reliable main effects of Block were observed from 300-800 ms (p = .001) and 1100-1400 ms (p = .001), reflecting improved Target searching in Block 2. No other clusters of main effects or interactions were significant (all p's ≥ .1).

Comparison of Exp 1A and 1B Results of a combined cluster-based analysis identified a main effect of Experiment from -100-1300 ms (p = .001), reflecting the sizable and extended advantage for those given preview of the images (Exp 1B). In addition, there was a main effect of the Patient Posture features of the Target Image from 500-900 ms, such that Target images in which the Patient had (congruent) Patient-like Posture received slightly more looks than Target images whose Patient had (incongruent) Agent-like Posture. There was also a main effect of Block from 400-1600 ms, reflecting the fact that Targets were identified faster overall in the second half of the experiment. The effect of Foil Patient Posture image interacted with Experiment, from 400-700 ms, reflecting that a Patient-like Posture provided a greater advantage for Target looks when images were previewed (Exp 1B) than not (Exp 1A). Finally, some three-way interactions were identified, but these are very likely under-powered tests and so will not be discussed further.

Discussion

Experiment 1 offered evidence that rapid gist extraction of event information is available to ongoing linguistic processes. Even when participants could not see the action images until the onset of the Verb in the sentence (Exp 1A), looks to the Target were reliably above chance 400-500 ms post Verb Onset. Accounting for time to execute a saccade, this means that visually-recognized event information was likely available to linguistic processing 300-400 ms post Verb Onset. Providing preview of the images (Exp 1B) moved above-chance Target looking 200 ms earlier (200-300 ms post Verb Onset), suggesting preview permitted earlier recognition of event information. Contrary to predictions, effects of Patient Posture (in Targets and Foils) had relatively late and weak effects on looking behavior. We expected an advantage when the images contained Patient-like Patients because this would speed event category classification for both images. However, only the Target Patient Posture image had the expected effect in the combined analysis.

Experiment 2

The lack of effects of Foil Patient Posture in Exp 1 may be the result of incremental linguistic processing. Hearing "The red person is…" likely focuses attention on the red person in both the Target and the Foil, making the Patient Posture (e.g. blue person) less relevant, especially in the Foil image because the Foil Agent's Posture does not match the verb's specific posture properties. i.e., although the Foil Agent has Agent-like properties - outstretched appendages, facing toward the other - his specific posture supports tapping, not kicking. Exp 2 addresses these issues.

Participants in Exp 2A viewed image pairs starting at Verb Onset and those in Exp 2B at Sentence Onset. Unlike Exp 1, the Foil image always had event roles opposite of the Target. If the Target had a red Agent, the Foil had a blue Agent, and

vice versa. This change should have a large effect on image preview, because in the Sentence Onset condition, hearing, e.g. "The red person is…" should permit identification of the Target because it has the red Agent, leading to above chance Target looking prior to Verb Onset. This change, however, makes the Foil Patient Posture relevant because the Foil Patient matches the color mentioned in the utterance ("red"). An Agent-like Patient might, therefore, lure looks to the Foil because it could be an Agent. Thus, Foil Patient Posture may show effects (Patient-like > Agent-like).

Also, unlike Exp 1, Target properties were not manipulated: all Targets had a prototypical Patient-like stance. Instead, our second (fully crossed) factor was Foil Action type: the action was either the Same as the Target (e.g., kicking vs. kicking) or Different (e.g., kicking vs. tapping). If language processing is strongly incremental, this manipulation should have little or no effect, since hearing "The red person is…" is sufficient to identify the Target. More likely, the Subject Noun Phrase is a probabilistic constraint, which will be weighed with the verb information, leading to effects of Action Type: it will be easier to find the Target when the Foil has a Different Action.

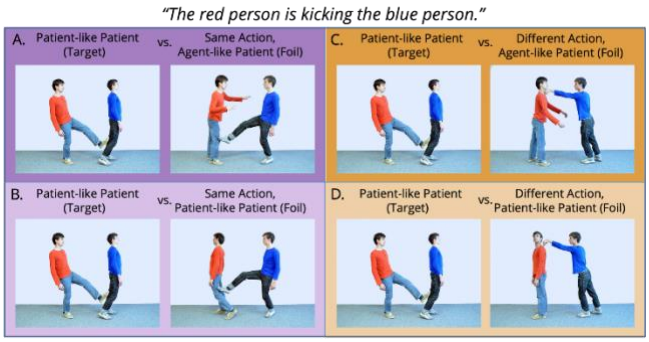

Figure 3: Example item, four conditions (Exp 2). Target image always had a Patient-like Patient. Foils had the Same Action as the Target (A and B) or a Different Action (C and D). Foils had an Agent-like Patient (A and C) or a Patient-like Patient (B and D). Color backgrounds not shown to participants but correspond to graphed conditions in Fig 4.

Methods

Participants One hundred and four native English speakers at the University of Pennsylvania participated for course credits. Forty participants from Experiment 2A and forty from 2B were included in the analyses. Twenty-four participants were excluded due to low clicking accuracy (7), being non-native speakers of English (7), computer errors (3), high track loss (2), repetitive lists (3), and failing to calibrate (2).

Materials The materials were the same Exp 1, except for the following. Target images always depicted actions with prototypical postures, while Foil images showed the same participants in reversed roles (see Fig3). The Foil images varied systematically along two dimensions in a 2×2 within-subject design: Action Type (whether the action was Different or Same as the Target) and Patient Posture (whether the Patient was in a Patient-like or Agent-like Posture).

Procedure Same as in Exp 1.

Analysis Same as in Exp 1.

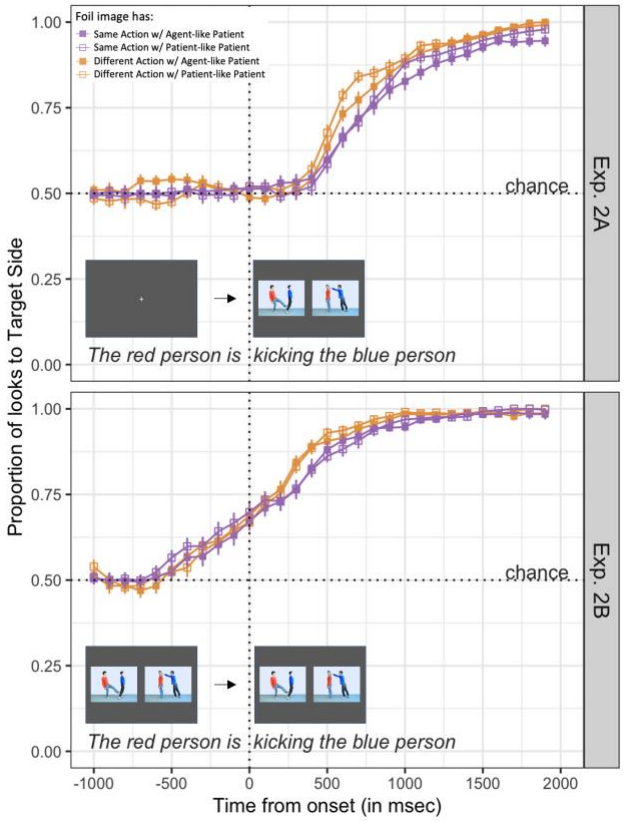

Figure 4: Proportion of looks to Target Side as a function of Time from Verb onset. Means and SE of subject means

Results Exp 2A (Images at Verb Onset)

Eye gaze was analyzed for correct trials only (98% of trials). Fig4 (top panel) plots the proportion of looks to the Target, relative to Verb Onset. Descriptively, participants show increased Target looks at about 400-500 ms. It appears that Target looks rise more when the Foil image had a Different action compared to a Same action, especially when Foil Patient had a Patient-like Posture rather than an Agent-like Posture.

Time to Access Event Information Using the same statistical method as Exp 1, we find that above-chance Target looking begins during the 400-500 ms sample (Est. = 0.227, t = 2.37, p = 0.032). Assuming a 100 ms lag for eye movements to respond to linguistic input, event recognition processes made contact with linguistic processes 300-400 ms into image (and verb) processing.

Cluster-based Permutation Tests The results (Table 2A) revealed a significant effect of the Foil Action Type occurred from 500-1400 ms (p = .006), reflecting greater looks to the Target if the Foil action was Different from the Target action. The potential interaction between this effect and Foil Patient Posture seen in Fig4 (top panel) was not significant (p > .1). A reliable main effect of Block was observed from 400-1600 ms (p = .001), reflecting improved Target searching in Block 2. No other clusters of main effects or interactions were significant (all p's > .1).

Table 2: Exp 2 cluster-based permutation tests results.

A.Effects (Exp2A)	Cluster in ms	Sum t	p =
Foil Action Type	500 to 1400	26.05	0.006
Block (1 vs. 2)	400 to 1600	-43.38	0.001
B.Effects (Exp2B)	Time of Cluster	Sum t	p =
Foil Action Type	300 to 1200	26.47	0.001
Block (1 vs. 2)	-200 to 1200	-62.95	0.001
C.Effects (Combined Exp2A and Exp2B)	Time of Cluster	Sum t	p =
Experiment (2A vs. 2B)	-500 to 1500	150.04	0.001
Foil Action Type	300 to 1300	35.49	0.001
Foil Patient Posture	700 to 1100	-9.02	0.001
Foil Action Type x Foil Patient Post.	500 to 700	-5.40	0.001
Foil Action Type x Experiment	600 to 800	-6.10	0.001
	1700 to 1900	-6.70	0.001
Block (1 vs. 2)	-300 to 1500	-76.13	0.001
Block x Foil Patient Posture	1200 to 1400	-5.89	0.001
(No other reliable effects or interactions)			

Formula (Exp 2A&2B): Elog ~ 1 + FoilAction*FoilPosture*Block + (1 | Subject) + (1|TargetVerb)
Formula (Combined): Elog ~ 1 + Exp*FoilAction*FoilPatient*Block + (1| Subject) + (1|TargetVerb)
Applied to 100 bins from -1000 to 2000. Significant clusters identified with jlmerclusterperm

Results Exp 2B (Images at Sentence Onset)

Time to Access Event Information Eye gaze was analyzed for correct trials only (97% of trials). Using the same statistical method as Exp 1, we find that above chance Target looking occurred at Verb Onset (Est. = 0.946, t = 11.1, p < 0.0001).

Cluster-based Permutation Tests The results (Table 2B) revealed a significant effect of the Foil Action Type from 300-1200 ms (p = .001), reflecting greater looks to the Target if the Foil action was Different from the Target action. A reliable main effect of Block was observed from -200-1200 ms (p = .001), reflecting improved Target searching in Block 2. No other clusters of main effects or interactions were significant (all p's > .1).

Comparing Exps 2A and 2B The combined analysis revealed a reliable effect of Experiment (2A vs. 2B) from -500-1500 ms (p = .001), reflecting substantial anticipatory Target looking when the images appeared at Sentence Onset. There was a reliable effect of Foil Action Type (Different > Same) from 300-1300 ms (p = .001). This effect of Action Type was larger when there was no preview (Exp 2A), resulting in a reliable interaction between Foil Action Type and Experiment from 600-800 ms (p = .001) and 1700-1900 ms (p = .001). And while the effect of Foil Patient Posture did not reach significance in either experiment individually, a reliable main effect of Foil Patient Posture appeared 700-1100 ms (p = .001), reflecting greater looks to the Target when the Foil Patient had a Patient-like Stance. Although the apparent interaction between Action Type and Foil Patient Posture was not significant separately in Exp 2A and 2B, it was significant in the combined analysis from 500-700 ms (p = .001). Finally, there was a reliable effect of Block extending from -300-1500 ms, which interacted with Foil Patient Posture from 1200-1400 ms.

Discussion

Exp 2 offered evidence that rapid gist extraction of event information is available to ongoing linguistic processes. Even when participants could not see the action images until the onset of the Verb in the sentence (Exp 2A), their looks to the Target were reliably above chance 400-500 ms post Verb Onset. Thus, visually-recognized event information was likely available to linguistic processing 300-400 ms post Verb Onset. Providing preview of the images (Exp 2B) produced the predicted strong effect: since hearing 'The red person is...' provided sufficient information to identify the Target image, participants demonstrated above-chance Target looking even before hearing the verb. Contrary to what was predicted, an Agent-like Posture for the Foil did not lure looks away prior to the Verb. This did not happen until well after hearing the Verb across both Exps (800-1100 ms). Thus, a certain amount of non-incremental processing is observed in both Exps. We suggest this may be an act of refining event representations prior to responding, or simply double checking the Foil image prior to responding.

General Discussion

In two eyetracking experiments of linguistically guided visual search, we find evidence not only for rapid gist extraction of event information but also for its connection to ongoing linguistic processing. When participants were given no preview of the Target and Foil actions until hearing the Verb in the sentence, looks to the Target action were estimated to exceed chance 400-500 ms post-Verb Onset for both Exp 1A and 2A. This timing confirms our first hypothesis regarding rapid access to event information and aligns with prior work on Event gist extraction (Hafri et al., 2018).

Some evidence was also provided that postural information of the individuals in these images affected search in the expected directions. However, most effects were weak and arose somewhat later in the eye movement record than anticipated. We suggest that further visual interrogation of the scene allows not only for the refinement of event and event structure but also is used to verify Target choice in this task. If this is the case, postural information, particularly atypical postures, may be expected to be re-examined by observers.

References

Altmann, G. T. (2011). Language can mediate eye movement control within 100 milliseconds, regardless of whether there is anything to move the eyes to. *Acta Psychologica, 137*(2), 190-200.

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

De Freitas, J., & Hafri, A. (2024). Moral thin-slicing: Forming moral impressions from a brief glance. *Journal of Experimental Social Psychology, 112*, Article 104588. https://doi.org/10.1016/j.jesp.2023.104588

Dobel, C., Gumnior, H., Bölte, J., & Zwitserlood, P. (2007). Describing scenes hardly seen. *Acta Psychologica, 125*(2), 129-143.

Glanemann, R., Zwitserlood, P., Bölte, J., & Dobel, C. (2016). Rapid apprehension of the coherence of action scenes. *Psychonomic bulletin & review, 23*, 1566-1575.

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

Huettig, F., Rommers, J., & Meyer, A. S. (2011). Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta Psychologica, 137*(2), 151-171.

Ji, H., & Scholl, B. J. (2024). "Visual verbs": Dynamic event types are extracted spontaneously during visual perception. *Journal of Experimental Psychology: General, 153*(10), 2441.

Sun, Z., Ke, Q., Rahmani, H., Bennamoun, M., Wang, G., & Liu, J. (2022). Human action recognition from various data modalities: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 45*(3), 3200-3225.

Ullman, S. (1996). Visual cognition and visual routines. In *High-Level Vision: Object Recognition and Visual Cognition*, pp. 263–315, MIT Press, Cambridge, MA.

Vettori, S., Odin, C., Hochmann, J.-R., & Papeo, L. (2024). A perceptual cue-based mechanism for automatic assignment of thematic agent and patient roles. *Journal of Experimental Psychology: General.* Advance online publication. https://doi.org/10.1037/xge0001657