Are Abstract Relational Roles Encoded Visually? Evidence from Priming Effects

Alexander Alexandrov Petrov (apetrov@alexpetrov.com)
The Ohio State University, Department of Psychology
Columbus, OH 43210 USA

Yuhui Du (avaydu@gmail.com)
The Ohio State University, Department of Psychology
Columbus, OH 43210 USA

Abstract

It remains controversial whether the visual system encodes abstract relational roles such as *Agent* and *Patient* in visual events. The present experiment tested whether abstract role bindings induce priming effects across consecutive events. Each trial included a static target image preceded either by a brief silent video of a priming event or by an audio-visual presentation of an English sentence describing the same event. Example sentence: "The red goat on the left knocked down the blue goat on the right." 64 videos counterbalanced 4 event types: launching, deforming, breaking, and a relationally ambiguous control. The set of static targets were the final frames of the same videos. The role bindings were either repeated, switched, or ambiguous across the target and prime. The dependent variable was the latency on a color-localization task (e.g., whether the red animal was on the left or on the right). Whereas the linguistic primes had no statistically significant effect on the latency of the visual task, the role bindings of the video primes did have an effect: The latency on unambiguous trials (which required role binding) was significantly greater than that on ambiguous trials (on which at least one component lacked clear relational roles). This suggests the visual system is sensitive to (the ambiguity of) the role bindings of abstract relations.

Keywords: visual relational processing, abstract event roles, priming

Visual Processing of Abstract Relations?

Relational processing plays a crucial role in human cognition. The ability to represent and process abstract relations is indispensable for high-level cognitive functions such as reasoning, analogy-making, and language (e.g., Gentner & Markman, 1997; Hummel & Holyoak, 1997; Jackendoff, 2003). One important class are *eventive relations* – who did what to whom. Linguists analyze them in terms of *thematic roles* such as Agent, Patient, Instrument, Location, etc. (e.g., Bornkessel, Schlesewsky, Comrie, & Friederici, 2006; Fillmore, 1968). Many cognitive scientists posit (or often simply presuppose) the existence of amodal conceptual representations of objects, events, and relations. Influential examples of such amodal representational frameworks are Fodor's (1975) *Language of thought*, Newell & Simon's (1976) *Physical symbol systems*, and the declarative structures in Anderson's (1983, 2007) *ACT-R* architecture. In this article we take for granted the existence of some form of amodal relational representations.

Our motivating question here is whether abstract relational roles such as Agent and Patient can be *seen* directly in addition to being *thought of* and *talked about*. Obviously, when we look at a scene we become aware of who does what to whom. This indicates that the visual system ultimately constructs amodal conceptual representations. This is how we can think of and talk about what we see (Jackendoff, 2003). But there is a wide gap between low-level visual elements such as textures, surfaces, and motion patterns, on the one hand, and abstract Agents and Patients, on the other. The visual system consists of multiple processing stages (e.g., Felleman & Van Essen, 1991). High-level visual processing bridges the gap (Cavanagh, 2011, 2021; Tversky & Zacks, 2013; Ullman, 1984).

Our working hypothesis is that there exists a level in the visual hierarchy (presumably near the top) that constructs relational representations that are modality-specific enough to be labelled "visual" and yet generalize over enough low-level attributes to be labelled "abstract." We focus on the Agent and Patient roles in simple events.

The visual system represents and processes magnitude comparisons such as *larger*, *brighter*, and *more numerous* (e.g., Michal, Uttal, Shah, & Franconeri, 2016). Also, it represents and processes spatial relations such as *above* and *inside* (e.g., Clevenger & Hummel, 2014; Franconeri et al., 2012; Logan, 1994). Hardly anyone disputes this. What is more controversial is whether the list extends to the physical (e.g., *hang*, *support*), social (e.g., *chase*, *meet*), and eventive (e.g., *pull*, *break*) domains. In a recent review Hafri and Firestone (2021) argued that the visual system does represent and process relations in all these domains. (See also the earlier review by Scholl and Tremoulet, 2000.) These authors proposed several "signatures" such as speed and automaticity that in their opinion demarcate the boundary between perception and cognition.

The present experiment is inspired by two published studies of the perception of Agent and Patient thematic roles (Hafri, Papafragou, & Trueswell, 2013; Hafri, Trueswell, & Strickland, 2018). In these studies, a series of photographs of two-person scenes were presented in rapid succession. In one experiment, the participants had to indicate whether the person in red shirt was on the left- or right-hand side of the photo (Hafri et al., 2018). The key result was that, although relational role was orthogonal to shirt color and was never explicitly mentioned, participants responded more slowly when the target's role switched from trial to trial (e.g., the red-shirted person went from being the Agent to being the Patient). Hafri and colleagues relied on two "signatures" in

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

their interpretation of these data. They argued that the rapid serial presentation forced speedy and automatic processing, which they assumed occurs only in perceptual systems.

On the other hand, there is convincing evidence from visual-search experiments that visual relational processing requires attention. The search for a relationally defined target in a crowded display is difficult and its latency scales linearly with the number of distractors (Logan, 1994, 1995). Furthermore, attention is required for integrating the parts of a single object (e.g., Stankiewicz, Hummel, & Cooper, 1998; Stankiewicz & Hummel, 2002) and for binding the elements of simple dynamic patterns such as biological motion (Cavanagh, Labianca, & Thornton, 2001). Indeed, according to the influential Feature Integration Theory, attention is required to bind feature-map representations into coherent objects organized in scenes (Treisman & Gelade, 1980; Treisman, 1998; Wolfe, 2012). More recent theories build upon Treisman's foundations and differentiate several distinct functions of visual attention including binding and indexing (Rensink, 2013). Both functions seem necessary not just for visual processing but also for *any* processing of relations.

A form of binding that is key to relational processing is *role-filler binding* (Burwick, 2014; Feldman, 2013). To illustrate, compare the events described by these sentences:

The goat pushed away the horse.

The horse pushed away the goat.

These events differ only in the binding (or *assignment*) of objects to relational roles. Clearly, this is a crucial difference! Various neural-network models of role-filler binding have been proposed (e.g., Hummel & Holyoak, 1997; O'Reilly, Busby, & Soto, 2003; Smolensky & Tesar, 2006).

The present experiment builds upon the pioneering work of Hafri and colleagues (2013, 2018) and develops it in light of theoretical considerations about the importance of attention in visual relational processing and the importance of role-filler binding in relational processing of any kind. Our approach extends the previous studies in three key aspects:

1. In addition to events with unambiguous assignment of objects to relational roles, our experimental design includes events with ill-defined roles and ambiguous assignment. The unambiguous events require role-filler binding for a proper representation, whereas no such binding is necessary to represent the ambiguous events. We hypothesize that these different information-processing requirements might lead to measurable differences in response times.

2. We used videos instead of static images to induce priming effects. We hypothesize that the magnitude of these effects will be greater than the 6-ms effect induced by the rapid serial presentation of Hafri et al. (2018). The introduction of separate priming-inducing stimuli sets the stage for the most important difference between our approach and previous work, namely:

3. We explicitly compare the effects of linguistic primes (English sentences) to video primes. Psycholinguistic studies have confirmed that thematic roles (including Agent and Patient) can prime subsequent utterances (e.g., Chang, Bock, & Goldberg, 2003; Hare & Goldberg, 1999). We aim to replicate Hafri et al.'s (2018) finding that the role assignment in one image can prime the role assignment in a subsequent image. Critically, however, our design allows us to measure cross-modal priming. Our hypothesis is that visual primes will have a measurable effect on subsequent visual relational processing, whereas linguistic primes will have no effect. In our opinion, such pattern would be much better evidence for the visual locus of the relational representation. This obviates the need to postulate "signature" properties of perception.

Experiment Method

Each trial of the main task consisted of a priming event followed by a static target image. The priming event was either presented as a video or described in an English sentence. Each stimulus involved a pair of animals: one red and one blue. Each participant was instructed to localize their assigned color in the static image: "Press A if the red animal is on the left and L if it is on the right." There were also demo trials and threshold-estimation trials (described later).

Stimuli and Apparatus

Sixty-four event templates generated 64 short videos and 64 corresponding sentences. For example, one sentence was: "The red lion on the left broke the blue lion on the right." The two animals in an event were always of the same species. The full factorial design crossed 4 species (goat, horse, lion, and mouse), 2 colors (agent is red vs. blue), 2 locations (red on the left vs. right), and 4 event types (Launch, Break, Deform, and Control).

Launch videos were inspired by Michotte's (1946) classic studies of the perception of causality. Agent always moved first and pushed Patient from the periphery of the "stage" towards the center. Neither animal changed shape at any time. They moved together for a few hundred milliseconds, then Agent stopped while Patient kept moving a little further. When Patient finally stopped, the two animals ended up in a mirror-symmetric configuration on the final video frame as illustrated on the top panel of Figure 1. When the final frame was viewed as a static image by itself, this symmetry made it ambiguous with respect to relational-role assignment.

The pattern of Agent's motion in Break and Deform videos was identical to that in Launch videos. The three event types differed only regarding Patient – it broke in half when the Agent impacted it in Break events, and changed shape while staying in one piece in Deform events. Thus, the final frames of Break and Deform videos contained enough visual cues to disambiguate the relational-role assignment and event type even when viewed in isolation as static images (Figure 1).

Control videos were designed to lack salient relational roles. The two animals "rocked" gently back and forth a few times in a mirror-symmetric manner without ever touching each other. Importantly, the final frames of Control videos were identical to the final frames of Launch videos. These two event types involved easily distinguishable patterns of motion. However, both motions stopped in the same ambiguous configuration with ill-defined roles (Fig. 1, top).

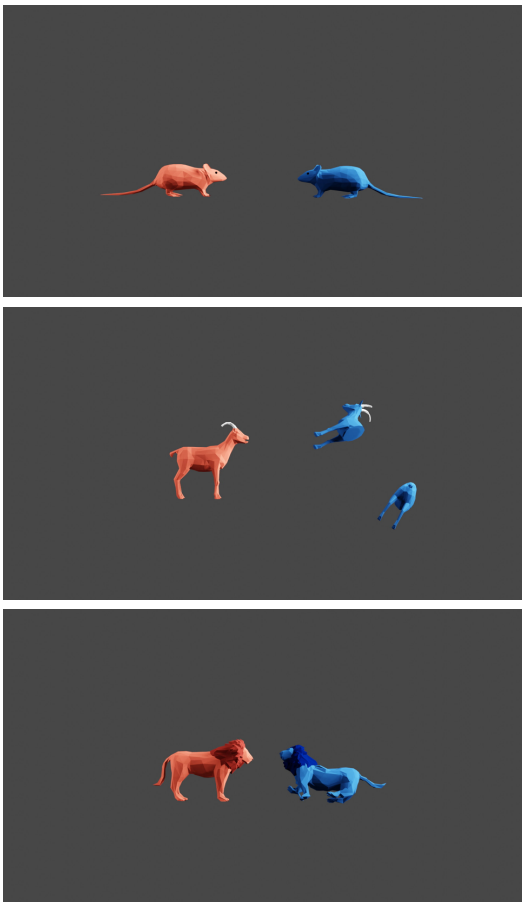

Figure 1: Examples of static images of 3 event types: Control (top), Break (middle), and Deform (bottom). The top image has ambiguous relational roles. The other two are unambiguous – Agent is red and on the left, Patient is blue and on the right. The same images are also the final frames of videos of these events. The fourth event type, Launch, is indistinguishable from Control in a static image (top panel).

The set of target images was the set of final frames in the event videos. Because Launch and Control events ended in the same configuration, the 64 videos generated only 48 distinct images, 3 of which are illustrated in Figure 1.

The 64 English sentences were designed to be as analogous as possible to the 64 videos. Launch events were described using the verb "pushed away." For instance, "The blue horse on the right pushed away the red horse on the left." The verb "broke" was used for Break events and "knocked down" for Deform events. Descriptions of Control events avoided transitive verbs altogether: "There was a blue lion on the left and a red lion on the right." All sentences were presented both as text written on the screen and as a male voice that played via the computer sound system. The synthetic audio files were generated with an online text-to-speech tool (https://www.readthewords.com/Try.aspx). All soundtracks were ≈3000 ms long. (Tracks of shorter sentences were padded with initial silence.) The videos were generated in an open-source animation software called *Blender* (Blender Foundation, 2022). Each video was 1400 ms in duration.

The experimental software was implemented in Javascript and was running on a server using the *psiTurk* framework (Gureckis et al., 2016). All data were collected in person on a client computer in a quiet darkened lab. The width of a target image spanned ≈15 degrees of visual angle at a viewing distance of ≈70 cm to the LCD monitor (refresh rate 60 Hz).

Experimental Design and Procedure

Each participant completed two sessions on separate days. Session 1 took 60 min on average and included instructions, demos, 100 threshold estimation trials, and 768 trials of the main task with video priming. Session 2 took 80 minutes on average and included a brief demo and 768 trials of the main task with linguistic priming. Each participant was assigned a *foreground color* (red or blue) in the beginning and their main task was to track this color throughout the entire experiment. This assignment was counterbalanced between participants.

Instructions and Demos: Participants viewed videos of a few representative events and were instructed that each event involved two animals of different colors, where one animal might "break", "knock down", or "push away" the other. The instructions mentioned explicitly that there were well-defined Agent and Patient for these three event types. Also, they explained that the animals didn't interact in "dance" events, in which case there were no clear Agent or Patient.

Threshold Estimation: Recall that some images contained enough visual cues to disambiguate the relational-role assignment. The purpose of the threshold estimation block was to estimate for each individual participant the minimal presentation duration that yielded 90% accuracy of Agent / Patient discrimination in such unambiguous cases. A set of 16 static images of type Break and 16 images of type Deform were sampled with replacement on each trial. No videos or ambiguous images were involved. The key sentence in the instruction was: "As quickly and accurately as possible, press A if the agent appears on the left and L if it appears on the right." The presentation duration was manipulated adaptively according to the staircase method (Leek, 2001). Concretely, we ran a 3-up-1-down staircase chain for 50 trials interleaved with a 4-up-1-down chain for 50 trials. That is, the duration was decreased after 3 (or 4, respectively) consecutive correct responses and increased after a single incorrect response. The range of possible durations was from 16.67 ms to 300 ms with a step size of 16.67 ms. Upon completion of the block, a transformed logistic function was fitted to the 100 responses and their corresponding durations using maximum likelihood (Shen & Richards, 2012). The participant's individualized threshold was set to the 90% point of their psychometric function (modulo the 60 Hz refresh rate of the monitor).

From this point on the instructions never mentioned the words "agent" or "patient" again. The participants were instructed to attend to color (rather than relational role). The first block of the main task began with a few practice trials to aid the transition to the color localization task.

Visual Priming: Each trial of the main task in the video session (Day 1) presented a video followed by a target image.

The participants were instructed to attend to the video because they might be asked a question about it at the end of the trial. However, their main task was to focus on the target image and press A if the animal of their assigned color was on the left or press L if it was on the right. Catch questions were asked on 72 of the 768 trials, chosen at random. After the participant keyed in their response on the primary task, the question appeared: "What happened in the short video?" The response options were "Break", "Knock down", "Push away", and "Dance/NA". There was no feedback. Some participants were excluded from the sample because of their unsatisfactory accuracy on these catch questions.

Each trial began with a video presentation for 1400 ms. The final frame of the video stayed on screen for an additional 100 ms. Then a pattern mask with a fixation cross appeared for 100 ms. The mask was a jumble of overlapping blue and red fragments of animals of various species. The target image was presented for the participant's individualized duration, followed by a different pattern mask. The mask remained on screen while the computer waited for a key press. If the response time was longer than 2000 ms, a "Slow Response" message appeared. No other feedback was given. After a 500 ms inter-trial interval, the next video started, etc.

Because pairing each of the 64 videos with each of the 48 target images generated too many combinations, each participant was exposed only to 1/4 of the design space by sampling only 4 of the 16 possible animal-animal pairings. We developed a counterbalanced scheme of 6 sampling templates and randomly assigned each participant to one of them. Each template was constrained so that the animals repeated from video to image on half of the trials, whereas on the other half the video featured one species and the image featured a different species. (Recall that the two animals in any given event *always* were of the same species.) For example, mouse videos would be paired with mouse images, lion videos with horse images and vice versa, and goat videos with goat images for a particular participant.

In this way, each of the 64 videos was paired with 12 target images. The template prescribed the image animal as a function of the video animal. The other image attributes were counterbalanced: 3 event types (Control, Break, and Deform; cf. Figure 1) by 2 relational-role assignments (agent is red vs. blue) by 2 locations (red on left vs right). The sequence of 768 visual priming trials was divided into 12 blocks of 64 trials in random order under the constraint that each video appeared exactly once in each block. The participants were encouraged (but not required) to rest between blocks.

Linguistic Priming: The linguistic session on Day 2 was completely analogous to the video session except that the priming stimuli were English sentences instead of videos. The session began with a few practice trials to remind participants of the color localization task. There were 768 linguistic priming trials organized into 12 blocks according to the same factorial design but in freshly randomized order. Catch questions on 72 trials asked what event was described in the sentence. The presentation sequence was analogous to that in the visual session, except that each trial began with a 3000 ms audio-visual presentation of a sentence instead of a video. The target images were displayed with the same timing and masking parameters as on Day 1.

Participants and Exclusion Criteria

The participants were students at the Ohio State University (OSU). Some of them participated for course credit, others as a favor to the experimenters. They came to the lab in person and gave informed consent as approved by OSU Institutional Review Board. They received $15 at the end of the second session. Twenty-three participants were recruited but two of them did not come back on Day 2. Six more were removed from the sample according to our pre-determined exclusion criteria. Specifically, 3 people were excluded for having chance-level accuracy (49%) on at least one session. One person had chance-level accuracy (26%) on the catch questions. Two others had relatively low accuracy on the catch questions (47% and 61%) and *also* on the main trials (87%). Overall, 15 participants remained for the main analysis. Eight of them searched for red and 7 for blue targets.

Data Preprocessing: Learning-Curve Estimation

One technical challenge in designing this experiment was that the audio presentation of a sentence lasted 1.6 seconds longer than a video. This added 20 extra minutes to the linguistic session and made it impractical to counterbalance the two stimulus modalities across days. Unfortunately, the resulting imbalanced design confounds the cross-modal manipulation with the effects of practice. In our study, the practice effects are controlled statistically. We estimated the response-time (RT) *learning curve* individually for each participant in each session. Concretely, both a linear and an exponential model were fitted to a given data segment (Heathcote, Brown, & Mewhort, 2000; Petrov & Hayes, 2010, Eq. 4). We used Akaike's information criterion for model selection (Burnham & Anderson, 2002, p. 63). The differences between the observed RTs and the trial-by-trial regression predictions – the *residual RTs* – do not depend systematically on time.

Results

The 15 participants who took the experiment seriously achieved very high accuracy on the main task: group mean = 96.7% (SD = 2.8% across participants) in the video and 98.0% (SD = 1.4%) in the linguistic session. The accuracy on the catch questions was also high: group mean = 82.3% (SD = 10.1%) in the video and 98.9% (SD = 1.3%) in the linguistic session. The latter difference is highly statistically significant ($t(14) = -6.13$, $p < 0.001$). The superior memory in the linguistic condition is easy to explain. The sentence could be stored in verbal working memory where it was sheltered from interference from the intervening target image. What is more important for our purposes is that the videos were encoded well enough to support relatively accurate recall of the event type two seconds later despite the visual interference.

The presentation-duration thresholds varied considerably across individuals (M = 193 ms, SD = 88 ms). This justifies the time and effort for threshold estimation on Day 1.

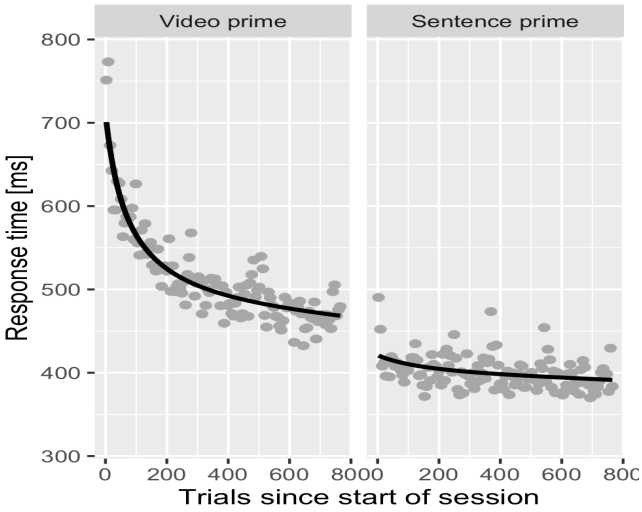

Figure 2: The learning curves in the two sessions (averaged across 15 individual curves). Each data point is the mean of 90 observations (15 participants x 6 trials).

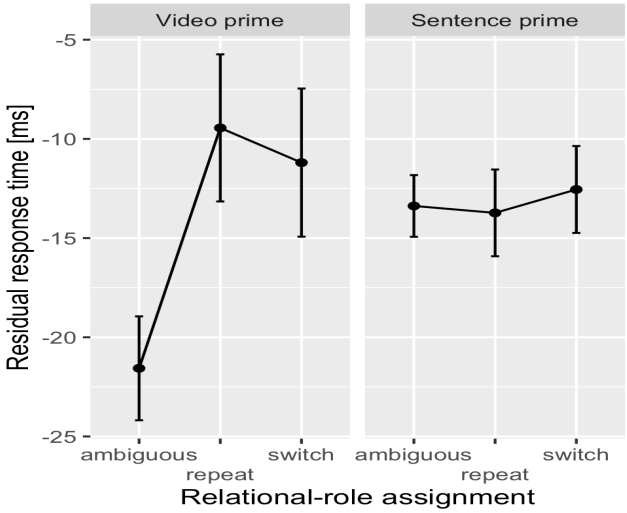

Figure 3: Group-averaged latency of the color localization task as a function of the role assignment of the preceding priming event. Error bars are 95% within-subject CIs.

We preprocessed the RT data by fitting learning curves as outlined above. There were large individual differences, as expected. In the video session the standard deviation of the mean latency was SD = 123 ms. The grand mean was 500 ms. The exponential model was selected for 11 of the individual learning curves (Figure 2, left). In the linguistic session the mean latencies were less variable (SD = 65.9 ms) and faster overall (grand M = 386 ms). Three learning curves were exponential and 12 were linear (and quite flat: Fig. 2, right).

The next data-processing step is of crucial methodological importance. We calculated the residual deviations from the gradual learning trends. Intuitively, this transformed the sloping curves in Figure 2 into a flat baseline that is identical for both sessions, thereby making cross-modal comparisons possible. It also eliminated the individual differences in mean latency. The dependent variable for all subsequent analyses is this *residual RT* (or *RRT* for short, measured in ms).

Next, we applied the outlier exclusion criteria of our pre-determined plan. We dropped all trials with incorrect responses and trials whose latencies were too fast (RT < 200 ms) or too slow (> 2000 ms from the onset of the target image). Finally, we also excluded trials whose residual RTs were more than 3 standard deviations away from the mean. A total of 4.6% of the trials were excluded due to all criteria combined. This left 21,973 data points for the main analysis.

Recall from the introduction that the variable of greatest theoretical interest is the assignment (or binding) of objects to relational roles. To illustrate, the red lion in the bottom panel of Figure 1 is bound to the Agent role, whereas the blue lion is bound to the Patient role. This is *unambiguous* role assignment. By contrast, the top panel illustrates *ambiguous* assignment – both animals have equal claim to either role. When *both* the prime (video or sentence) *and* the target image on a trial have unambiguous assignments, we say that the trial as a whole is unambiguous. This occurs when a prime of type Launch, Break, or Deform is paired with a target of type Break or Deform. Otherwise, the trial is ambiguous. This occurs when the prime or target (or both) are of type Control.

Unambiguous trials have either *repeated* (or *congruent*) or *switched* (or *incongruent*) role assignment. The former occurs when the Agent in the prime has the same color as the Agent in the target. The latter occurs when they have different colors. To express these distinctions in the statistical analyses we defined a Role-assignment factor with three levels: *ambiguous*, *repeated*, and *switched*.

For our main analysis we used a linear mixed-effect model that included Session (`S`) and Role-assignment (`R`) and their interaction as fixed effects, and Participant (`P`) as random effect. Concretely, we used the `lme4` package in R (Bates, Mächler, Bolker, & Walker, 2015; Kuznetsova, Brockhoff, & Christensen, 2017). The model was specified by the formula

```
RRT ~ S + R + S:R + (1+S | P)
```

The random-effect term `(1+S|P)` was included to eliminate the statistical artifacts introduced by the uneven exclusion of outliers across participants. We also fitted simpler models, but they were statistically inferior to the one above. All pointed to identical substantive conclusions.

The main effect of Session was included in the model as an explicit check whether we had succeeded in regressing out the learning effects in Figure 2. Indeed, the associated F statistic was close to zero ($F(1, 15) = 0.08$, $p = 0.78$). This verifies that we had established a common baseline for the *residual* RTs. (The *raw* RTs were faster in the linguistic session – notice the drop across the two panels of Figure 2.) This common baseline is slightly negative (–14.7) because most of the excluded outliers were "long." This asymmetry reflects the well-known positive skew of RT distributions.

The common baseline sets the stage for testing our working hypothesis – namely, that the RRTs would be affected by the Role assignment factor in the video session but not in the linguistic session. The RRT profiles in Figure 3 are entirely consistent with this prediction. Our most important result is

that the role bindings of the video primes did affect the latency of the subsequent color localization task. It was ≈11 ms slower on unambiguous than ambiguous trials (left panel). This difference is statistically highly significant ($t(15) = 5.9$, $p < 10^{-8}$) and indicates a priming effect. As predicted, priming was observed only in the video session (S-by-R interaction, $F(2, 21943) = 12.8$, $p < 10^{-5}$). The role assignments of linguistic primes had no statistically significant effect, again as predicted ($F(2, 11061) = 0.36$, $p = 0.6$). The tight error bars indicate considerable statistical power despite the relatively small number of participants. Thus, the null effect of linguistic primes can be interpreted as evidence of absence of priming, rather than as mere absence of evidence.

Finally, we performed a planned-comparison test of the contrast between repeated vs. switched role assignments on relationally unambiguous trials in the video session. There was no evidence of switch costs ($t(10882) = 0.65$, $p = 0.51$). We thus failed to replicate the small (6 ms) but statistically significant switch cost observed in a related experimental paradigm (Hafri et al., 2018).

Du (2023) reports further analyses involving other independent variables such as event type and visual location.

Discussion

We presented experimental evidence that the role bindings of video primes can affect the latency of a subsequent color localization task. The priming effect was sensitive to the ambiguity of the role assignments in the video and/or the target image. Concretely, the residual response times were ≈11 ms slower in the unambiguous condition compared to the ambiguous control. One plausible explanation of this pattern is that role-filler binding was required for a complete representation of the unambiguous stimuli, whereas the ambiguous cases called for less binding. If we assume that it takes time for the visual system to establish and maintain role-filler bindings, one expects slower responses in the unambiguous condition.

Furthermore, the priming effect was modality specific. As predicted, the linguistic primes had no significant effect on the latency of the subsequent visual task. This suggests that the Agent / Patient distinction that drives the priming effect is represented at a visual site rather than an amodal one.

This leads to the question of where the visual system ends and central cognition begins. This question is entangled in the literature with the thorny issue of modularity of visual perception (Firestone & Scholl, 2016; Pylyshyn, 1999). The lack of cross-modal priming in our study can be interpreted as evidence for domain specificity and information encapsulation in the sense of Fodor (1983). However, it is also compatible with a more gradualist conception that denies the existence of a sharp boundary between the visual system and the rest of the cognitive architecture. As stated in the introduction, our working hypothesis is that there exists a level in the visual hierarchy (presumably near the top) that constructs relational representations that are modality-specific enough to be *labelled* "visual." Our data are consistent with this gradualist interpretation.

The other half of our working hypothesis is that said relational representations generalize over enough low-level attributes to be *labelled* "abstract." This leads to another thorny question – that of the nature of the distinction between abstractness and concreteness (Campbell, 1990). One critical commentary to Hafri and Firestone's (2021) review puts the question bluntly: "How can it be both abstract and perceptual?" (Hochmann & Papeo, 2021). On that view, the phrase "visual processing of abstract relations" is a contradiction in terms. Similar criticisms have dogged the entire line of research stemming from Michotte's (1946) work. Do his launching displays convey true causality or something else? This echoes an age-old philosophical debate (e.g., De Pierris & Friedman, 2018).

The Agents in our stimuli varied in color, location, and animal species. The priming effect generalized across these three attributes. On the other hand, it must be acknowledged that the Agent in our videos always moved first and never changed shape. Conversely, the Patient moved second and did change shape. The shape changes were quite dramatic in Break events – the animal broke into two topologically disconnected pieces (Fig. 1, middle). The target images were static and thus conveyed no information about who moved first. Thus, the Patient's shape was the only feature that disambiguated the relational roles in (some of) the images. The Patient role was systematically confounded with certain shape deformations in our study. This raises the concern that our priming effects stem from concrete shape deformations rather than abstract thematic roles. The flat RRT profile in the linguistic condition (Fig. 3, right) indicates that this possible confound could not have operated on the static target images. We acknowledge that technically the confound had not been ruled out with respect to the low-level features in the videos. It is worth noting in this regard that in the experiments of Hafri et al. (2013, 2018), the relational roles were conveyed by the pose of the Agent rather than the Patient. For example, the Agents had outstretched arms and faced the Patient. Our study complements this earlier work by increasing the variety of stimulus materials that have been shown to induce priming effects. This variety makes it implausible that the effects might have been driven by low-level confounds.

Our design did not counterbalance the priming modality across days. Hence we had to go to great lengths to control statistically for the effects of practice. This methodology depends on certain assumptions – e.g., that priming- and practice effects combine additively. A better design would interleave video and linguistic blocks within both sessions.

In our future work, we plan to design a stimulus set in which no single visual feature is completely confounded with the relational role assignments across the board. For example, the Agent might have outstretched arms in some images, whereas the Patient might assume some characteristic pose in other images. The visual system must always rely on some disambiguating feature or other, but these features can change from one situation to the next. Although clairvoyance is impossible, the visual processing of abstract relations is not a contradiction in terms.

Acknowledgments

This article is based on the Ph. D. dissertation of the second author (Du, 2023, Chapter 4). We thank John E. Hummel for his insightful comments and support throughout this project.

References

Anderson, J. R. (1983). *The Architecture of Cognition*. Cambridge, MA: Harvard University Press.

Anderson, J. R. (2007). *How Can the Human Mind Occur in the Physical Universe?* New York: Oxford Univ. Press.

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effect models using `lme4`. *Journal of Statistical Software, 67* (1), 1-48.

Blender Foundation (2022). *Blender* (version 3.3.0) [Computer software]. https://www.blender.org

Bornkessel, I., Schlesewsky, M., Comrie, B. & Friederici, A. D. (Eds.) (2006). *Semantic Role Universals and Argument Linking*. Berlin: Mouton de Gruyter.

Burnham, K. P. & Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. New York: Springer-Verlag.

Burwick, T. (2014). The binding problem. *WIREs Cognitive Science, 5*, 305-315.

Campbell, K. (1990). *Abstract Particulars*. London: Oxford Blackwell.

Cavanagh, P. (2011). Visual cognition. *Vision Research, 51*, 1538-1551.

Cavanagh, P. (2021). The language of vision. *Perception, 50* (3), 195-215.

Chang, F., Bock, K., & Goldberg, A. E. (2003). Can thematic roles leave traces of their places? *Cognition, 90* (1), 29-49.

Clevenger, P. E. & Hummel, J. E. (2014). Working memory for relations among objects. *Attention, Perception, & Psychophysics, 76*, 1933-1953.

De Pierris, G. & Friedman, M. (2018). Kant and Hume on causality. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2018 Edition).

Du, Y. (2023). *An experimental investigation of relational* categorization and visual relational processing (Publication No. OSU1692626792393214) [Doctoral dissertation, The Ohio State University], OhioLINK Electronic Theses and Dissertations Center..

Feldman, J. (2013). The neural binding problem(s). *Cognitive Neurodynamics, 7*, 1-11.

Fillmore, C. J. (1968). The case for case. In E. Bach & R. T. Harms (Eds.), *Universals in Linguistic Theory*. New York: Holt, Rinehart, & Winston.

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

Fodor, J. A. (1975). *The Language of Thought*. New York: Thomas Y. Crowell.

Fodor, J. A. (1983). *Modularity of Mind: An Essay on Faculty Psychology*. MIT Press.

Gentner, D. & Markman, A. M. (1997). Structural alignment in analogy and similarity. *American Psychologist, 52*, 45-56.

Gureckis, T. M., Martin, J., McDonnell, J., Rich, A. S., Markant, D., Coenen, A., … & Chan, P. (2016). psiTurk: An open-source framework for conducting replicable behavioral experiments online. *Behavior Research Methods*, 829-842.

Feldman, J. (2013). The neural binding problem(s). *Cognitive Neurodynamics, 7,* 1-11.

Felleman, D. J. & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex, 1* (1), 1-47.

Franconeri, S. L., Scimeca, J. M., Roth, J. C., Helseth, S. A., & Kuhn, L. E. (2012). Flexible visual processing of spatial relationships. *Cognition, 122*, 210-227.

Hafri, A. & Firestone, C. (2021). The perception of relations. *Trends in Cognitive Sciences, 25* (6), 475-491.

Hafri, A., Papafragou, A., & Trueswell, J. C. (2013). Getting the gist of events: Recognition of two-participant action from brief displays. *Journal of Experimental Psychology: General, 142* (3), 880-905.

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.

Hare, M. L. & Goldberg, A. E. (1999). Structural priming: Purely syntactic? In M. Hahn & S. C. Stones (Eds.), *Proceedings of the 21st Annual Meeting of the Cognitive Science Society*. Mahwah, NJ: Erlbaum

Heathcote, A. J., Brown, S. D., & Mewhort, D. J. K. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic Bulletin & Review, 7*, 185-207.

Hochmann, J.-R. & Papeo, L. (2021). How can it be both abstract and perceptual? *Trends in Cognitive Sciences, 25*, https://doi.org/10.31234/osf.io/hm49p

Hummel, J. E. & Holyoak, K. J. (1997). Distributed representation of structure: A theory of analogical access and mapping. *Psychological Review, 104*, 427-466.

Jackendoff, R. (2003). *Foundations of Language: Brain, Meaning, Grammar, Evolution*. New York: Oxford University Press.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). `lmerTest` package: Tests in linear mixed effect models. *Journal of Statistical Software, 82* (13), 1-26.

Leek, M. R. (2001). Adaptive procedures in psychophysical research. *Perception & Psychophysics, 63* (8), 1279-1292.

Logan, G. D. (1994). Spatial attention and the apprehension of spatial relations. *Journal of Experimental Psychology: Human Perception and Performance, 20* (5), 1015-1036.

Logan, G. D. (1995). Linguistic and conceptual control of visual spatial attention. *Cognitive Psychology,28*, 103-174.

Michal, A. I., Uttal, D., Shah, P., & Franconeri, S. L. (2016). Visual routines for extracting magnitude relations. *Psychonomic Bulletin & Review, 23*, 1802-1809.

Michotte, A. (1946/ English transl. 1963). *The Perception of Causality*. Basic Books.

Newell & Simon, H. A. (1976). Computer science as empirical inquiry: Symbols and search. *Communications of the ACM, 19*, 113-126.

O'Reilly, R. C., Busby, R. S., & Soto, R. (2003). Three forms of binding and their neural substrates: Alternatives to temporal synchrony. In A. Cleeremans (Ed.), *The unity of consciousness: Binding, integration, and dissociation*. Oxford: Oxford University Press.

Petrov, A. A. & Hayes, T. R. (2010). Asymmetric transfer of perceptual learning of luminance- and contrast-modulated motion. *Journal of Vision, 10* (14):11, 1-22.

Pylyshyn, Z. (1999). Is vision continuous with cognition? The case for cognitive impenetrability of visual perception. *Behavioral and Brain Sciences, 22* (3), 341-365.

Rensink, R. A. (2013). Perception and attention. In D. Reisberg (Ed.), *The Oxford Handbook of Cognitive Psychology*. New York. Oxford University Press.

Scholl, B. J. & Tremoulet, P. D. (2000). Perceptual causality and animacy. *Trends in Cognitive Sciences, 4* (8), 299-308.

Shen, Y. & Richards, V. M. (2012). A maximum-likelihood procedure for estimating psychometric functions: Thresholds, slopes, and lapses of attention. *Journal of the Acoustical Society of America, 132* (2), 957-967.

Smolensky, P. & Tesar, B. (2006). Symbolic computation with activation patterns. In P. Smolensky & G. Legendre (Eds.), *The harmonic mind: From neural computation to optimality-theoretic grammar*. MIT Press.

Stankiewicz, B. J., Hummel, J. E., & Cooper, E. E. (1998). The role of attention in priming for left-right reflections of object images: Evidence for a dual representation of object shape. *Journal of Experimental Psychology: Human Perception and Performance, 24* (3), 732-744.

Stankiewicz, B. J. & Hummel, J. E. (2002). Automatic priming for translation- and scale-invariant representations of object shape. *Visual Cognition, 9* (6), 719-739.

Treisman, A. (1998). Feature binding, attention and object perception. *Proceedings of the Royal Society, London, B 353,* 1295-1306.

Treisman, A. & Gelade, G. (1980). A feature integration theory of attention. *Cognitive Psychology, 12*, 97-136.

Tversky B. & Zacks, J. M. (2013). Event perception. In D. Reisberg (Ed.), *The Oxford Handbook of Cognitive Psychology*. New York. Oxford University Press.

Ullman, S. (1984). Visual routines. *Cognition, 18*, 97-159.

Wolfe, J. M. (2012). Establishing the field: Treisman and Gelade (1980). In J. Wolfe & L. Robertson (Eds.), *From Perception to Consciousness: Searching with Anne Treisman*. New York: Oxford University Press.