Acta Psychologica 164 (2016) 157–164

Contents lists available at ScienceDirect

Acta Psychologica

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

# The pieces fit: Constituent structure and global coherence of visual narrative in RSVP

Carl Erick Hagmann [a,*], Neil Cohn [b]

[a] Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, United States
[b] Department of Cognitive Science, University of California, San Diego, La Jolla, CA, United States

ARTICLE INFO

*Article history:*
Received 16 April 2015
Received in revised form 19 January 2016
Accepted 20 January 2016
Available online xxxx

*Keywords:*
Visual language
Comics
Constituent structure
Visual cognition
Narrative
RSVP

ABSTRACT

Recent research has shown that comprehension of visual narrative relies on the ordering and timing of sequential images. Here we tested if rapidly presented 6-image long visual sequences could be understood as coherent narratives. Half of the sequences were correctly ordered and half had two of the four internal panels switched. Participants reported whether the sequence was correctly ordered and rated its coherence. Accuracy in detecting a switch increased when panels were presented for 1 s rather than 0.5 s. Doubling the duration of the first panel did not affect results. When two switched panels were further apart, order was discriminated more accurately and coherence ratings were low, revealing that a strong local adjacency effect influenced order and coherence judgments. Switched panels at constituent boundaries or within constituents were most disruptive to order discrimination, indicating that the preservation of constituent structure is critical to visual narrative grammar.



## 1. Introduction

Perceiving and integrating events over time is critical to the continuous dynamics of cognition (Corbetta & Shulman, 2002; Shipley & Zacks, 2008; Spivey & Dale, 2006). Humans and other animals can detect both static and dynamic changes in their environment over time (Hagmann & Cook, 2013; Rensink, 2004; Wright et al., 2010), but only humans can integrate information into a *narrative*, in which events depicted visually are interpreted as related and consequential to each other. Such comprehension is critical to understanding plots, stories, and instructions, and involves balancing a variety of covarying elements such as order, duration, and emphasis of component parts. While verbal narratives have been extensively researched, visual narratives have not, despite their prevalence in human culture for thousands of years—whether found on cave paintings, tapestries, or, in contemporary society, in the sequential images of comics (Kunzle, 1973; McCloud, 1994). Research into the comprehension of visual narratives has only recently emerged with seriousness and a focus on cognition (Cohn, 2013a; Magliano & Zacks, 2011; McCloud, 1994). We here explore one facet of this broader comprehension: the demands placed on perception and cognition in a rapidly presented visual narrative sequence.

Early research on sequential image understanding focused on the linear relations between images. Image-by-image comprehension involves continuously updating aspects of comprehension that rely on rapid scene understanding (Greene & Oliva, 2009; Potter, Wyble, Hagmann, & McCourt, 2014), and observing the changes that occur across characters, spatial location, and time (Magliano & Zacks, 2011; McCloud, 1994; Saraceni, 2001). Shifts in these dimensions (e.g., the introduction of a new character) incur costs in processing as a mental model of the narrative becomes updated with new information (Magliano, Dijkstra, & Zwaan, 1996; Magliano & Zacks, 2011; Zwaan & Radvansky, 1998).

Beyond these linear relations between images, *Visual Narrative Grammar* (VNG) argues that images in sequences take on narrative roles that are then combined into hierarchic constituents analogous to the way that sequential words take on syntactic roles that combine into constituents in sentences (Cohn, 2013b). This analogy is one at the functional level: a narrative grammar packages discourse-level meaning into a sequence using architectural constraints (categories, hierarchy, etc.) that are similar to the way that syntax packages meaning in sentences. VNG thus finds surface similarities with previous "grammatical" approaches to discourse (e.g., Clark, 1996; Hinds, 1976) particularly the well-known "story grammar" paradigms (e.g., Mandler & Johnson, 1977; Rumelhart, 1975), but differs from these precedents in both theoretical formalisms and the experimental methods used to provide evidence (see Cohn, 2013b for details). Experimentation has supported the idea that narrative structure in visual sequences is separate from its semantics (Cohn, Paczynski, Jackendoff, Holcomb, & Kuperberg, 2012), organized into constituents (Cohn, Jackendoff, Holcomb, & Kuperberg, 2014) and involves narrative categories defined by both content and context (Cohn, 2014).

* Corresponding author at: 426 Ostrom Ave, Syracuse, NY 13210, United States.
*E-mail address:* cehagman@syr.edu (C.E. Hagmann).

http://dx.doi.org/10.1016/j.actpsy.2016.01.011
0001-6918/

These structures can best be understood through an example. Fig. 1 illustrates a visual sequence with two constituents where Schroeder is playing in a sandbox with Snoopy. The sequence starts with an "Establisher" which sets up the situation of him playing in the sandbox. An "Initial" panel begins the events of the sequence, as Schroeder suddenly feels the heat of the sun. He then vigorously builds a sand mound in the subsequent "Peak" panel, a narrative climax of the primary actions of the sequence. The sequence is then resolved in the next panel, a "Release"—a resolution, aftermath, or coda of an action—where he rests in his newfound shade. A second constituent then begins suddenly with an even more climactic Peak, with Snoopy suddenly blowing the sand onto Schroeder, who then finds himself coated in the sequence-ending Release.

Importantly, narrative categories apply both to panels and to whole constituents. Together, the first four panels form their own constituent (an Initial) that, as a whole, sets up the entire second constituent (a Peak) at a higher level of structure. Each constituent is motivated internally by a Peak, which acts as the "head" of that constituent (double-barred lines). The penultimate panel of Snoopy blowing sand is thus the narrative climax of the whole sequence, reflected in its status as the Peak panel motivating the Peak constituent. The canonical *Establisher–Initial–Peak–Release* pattern is thus used in part or full at various levels of structure. These top–down global structures interact with the bottom-up content of images to determine the roles that images play in the sequence (Cohn, 2013b, 2014).

Initial evidence for the psychological reality of this narrative grammar came from experiments that used 1500 ms/panel sequences that balanced the contributions of narrative structure and/or semantic associative relationships across images (Cohn et al., 2012). Response times to panels in a target monitoring task were faster for panels in normal sequences, with both structure and meaning, than fully scrambled sequences of images with no narrative and no meaningful relations across images. However, intermediate response times resulted from target images in sequences with only semantic associations and no narrative structure (i.e., that maintained a common theme across panels) and from sequences with only narrative structure but no semantic associations (i.e., visual narrative analogs to a sentence like *Colorless green ideas sleep furiously*, which has syntax but no clear meaning). Across all sequences types, response times decreased across the ordinal position of sequences. Such results showed that narrative structure provides a behavioral advantage to the processing of sequences.

Another experiment using the same stimuli measured event-related brain potentials (Cohn et al., 2012), specifically the N400, a neural response typically lasting from 250 to 500 ms peaking around 400 ms, and thought to reflect the activation state of an incoming stimulus in semantic memory (Kutas & Federmeier, 2011). Panels from sequences with only semantic associations produced larger amplitude N400s than normal sequences. Even larger N400s appeared with scrambled and narrative-only sequences. Both scrambled and narrative-only sequences lacked coherent semantic associations between images, but the narrative-only sequences did have a felicitous narrative structure. Yet, because these amplitudes did not differ between scrambled sequences and narrative-only sequences, it confirmed that this narrative grammar was different from meaning, since the N400 was not attenuated by the presence of narrative structure. The N400 was, however, attenuated across ordinal panel position only in normal sequences, suggesting that a facilitation of meaning only occurs in the presence of both coherent narrative and semantic associations across images. Thus, while the low-level semantic information between images is involved in the comprehension of sequential images, it interacts with the global narrative structure.

Other studies have used techniques of rearranging images in visual narrative sequences to analyze their global structure and the roles taken by panels within a sequence. When viewing sequences at their own pace, comprehenders spend more time viewing panels from fully scrambled sequences than from coherent sequences (Cohn & Wittenberg, 2015; Foulsham, Wybrow & Cohn, submitted for publication). This slowing even occurs on the opening image of a sequence, where context has not yet rendered a sequence as incomprehensible, which suggests that some images are better candidates to open a narrative sequence than others (Cohn & Wittenberg, 2015; Cohn, 2014; Foulsham et al., submitted for publication). More targeted switching of panel positions within 4-panel sequences showed that comprehension worsens when panels are switched across distances than when switched locally (Cohn, 2014). Similarly, when participants were given four panels and asked to arrange them in an order that makes sense, misplaced panels were moved to adjacent positions more often than positions further in a sequence (Cohn, 2014). This *adjacency effect* was likely related to some images being more central to the narrative, and being surrounded by more peripheral images, which play more flexible roles in the sequence.

This global scope of narrative structure also must take into account the constituents formed by groupings of panels. Studies have long shown that participants are highly consistent in where they choose to divide both drawn and filmed visual sequences into sub-episodes (Cohn & Bender, submitted for publication; Gernsbacher, 1985; Magliano & Zacks, 2011). While research on this segmentation has typically viewed changes in linear coherence (such as shifts in characters or locations) as indicative of constituency boundaries (Gernsbacher, 1985, 1990; Magliano & Zacks, 2011; Radvansky & Zacks, 2014; Zacks, 2014), research within the VNG paradigm has shown that constituent structures go beyond such transient semantic changes. For example, narrative category information has been shown to be more predictive of conscious segmentation of drawn visual sequences than linear coherence relationships, though both do influence such divisions (Cohn & Bender, submitted for publication). Furthermore, measuring event-related potentials, Cohn et al. (2014) found that blank "disruption panels" placed within the constituents of visual narratives elicited a larger left anterior negativity than those placed between constituents, and this neural response was similar to those evoked by manipulations of grammar in language (Hagoort, 2003; Neville, Nicol, Barss, Forster, & Garrett, 1991). This effect could not be attributed to changes in linear coherence, because the amplitude to disruption panels was greater for

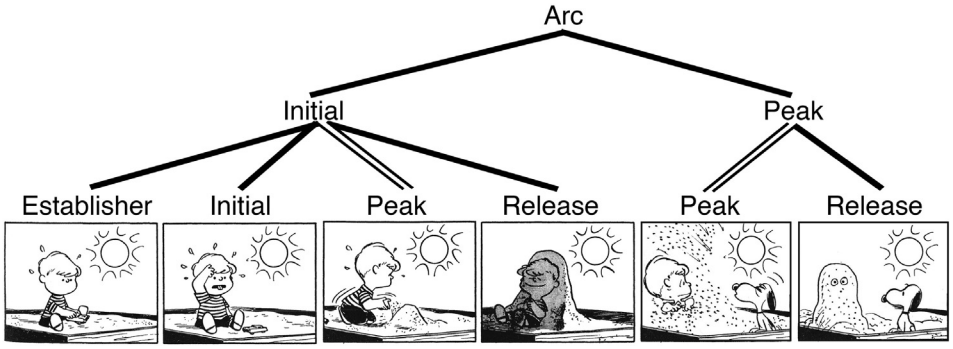


**Fig. 1.** Structure of a visual sequence with narrative categories and constituents.

those in the first constituent than those between constituents, in both cases *prior* to the crossing of the constituent boundary. Rather, this suggested a prediction of the upcoming constituent structure based on the content of the preceding panels.

These structure-specific constraints have been complemented by more general findings. For instance, the opening panel of a visual narrative sequence is viewed for longer durations than subsequent images in self-paced tasks, even in sequences with scrambled orders of images (Cohn & Wittenberg, 2015; Cohn, 2014; Foulsham et al., submitted for publication). First panel slowing is also consistent with studies of verbal discourse, in which slower reading is observed in the first sentence of a discourse (Glanzer, Fischer, & Dorfman, 1984; Haberlandt, 1984). Longer self-paced exposure to the first unit of a discourse is likely critical to "laying a foundation" for the subsequent structure (Gernsbacher, 1990), in which the comprehender accesses the primary elements of the situation, such as characters and location, before progressing to the subsequent narrative (Cohn & Paczynski, 2013; Gernsbacher, 1990). VNG has hypothesized that this general cognitive function is the purpose for the narrative category of "Establishers"—to devote a panel for this process of laying a foundation prior to the primary events of sequence (Cohn, 2013b).

Viewing time seems to be a major factor in visual narrative comprehension, but has been understudied relative to more specific structures of narrative and semantics. Previous experiments examining visual narratives with timed presentation typically have used panel durations at 1350 or 1500 ms (Cohn et al., 2014, 2012; West & Holcomb, 2002), but mean response times to target panels have been observed as low as 620 ms (Cohn et al., 2012), and self-paced studies have recorded mean viewing times for panels as short as 700 ms and as long as 2000 ms, depending on their content and ordinal sequence position (Cohn & Paczynski, 2013; Cohn & Wittenberg, 2015). However, some research has suggested that much shorter image durations can allow for visual sequence comprehension. Inui and Miyamoto (1981) used 4-panel comic strips with varying ISIs to explore viewers' sensitivity to switched panels. They switched the third panel with one of its adjacent panels on half the trials, and panel duration ranged from 83 ms to 150 ms between trials and experiments. Between each panel, a mask pattern was shown for a duration that ranged from 50 to 300 ms. Participants' ability to judge panel ordering was very poor with the briefest stimulus duration (83 ms with 300 ms ISI, 24% correct) and with the briefest ISI (150 ms stimulus duration with 50 ms ISI, 19% correct). In conditions with 150 ms stimulus duration and at least 100 ms ISI, participants performed very well (>62% accuracy).

In general, longer ISIs contributed to greater comprehension, and there was no significant effect of having a mask intersperse the panels instead of a blank. The authors compared their findings with those of Potter (1976), who showed that understanding briefly presented pictures relies on short-term conceptual memory. Interference with the encoding of these memories, through rapid serial visual presentation, could be the reason for better performance given longer processing time.

Recent explorations of the limits of visual conceptual understanding have used briefly presented pictures in ultrafast RSVP tasks (<80 ms/picture) (Potter & Hagmann, 2014; Potter et al., 2014). These studies have shown that the human visual system can extract meaningful information about a target picture among a sequence of pictures presented for only 13 ms each. Such rapid cognition allows for above-chance detection of the target picture within the sequence. Those studies required detection of a single picture among a sequence of many, instead of integration of narrative structure across the entire sequence. This raises the question: How exactly does a series of rapidly presented sequential images convey an understanding of a concept or narrative?

To answer this question, we tested the hypothesis that faster presentation rates of novel 6-panel long graphic sequences would reduce accuracy of discriminating ordered and mixed sequences and reduce overall coherence. We suspected that smaller distances between two switched panels would reduce order discrimination accuracy and promote coherence due to the adjacency effect facilitating local comprehension. In addition, our 6-panel long sequences allowed us to look beyond the simple structures found in prior studies of switched panels (Inui & Miyamoto, 1981; Cohn, 2014), and to analyze the effect of switching panels in sequences with distinct constituent structures. We hypothesized that switches between constituents, which violate both the groupings and the global structure of the sequence, would be more accurately recognized and less coherent than switched panels that remain within a constituent, which do not violate the global structure, yet do alter the local structure. Finally, we manipulated how long the initial panel was exposed to the participant relative to the remaining panels in a trial in order to test whether "laying a foundation" (Gernsbacher, 1990) is modulated by stimulus exposure.

Inui and Miyamoto (1981) found that detection of proper panel ordering completely falls apart with very brief exposure (<100 ms). Because our strips were longer (six instead of four panels), and had more possible types of panel switches (six instead of two), and because we sought to obtain coherence ratings from participants, we chose to examine discrimination between ordered and unordered sequences with panel durations of 1000 and 500 ms. With these relatively slow rates (compared to many RSVP experiments), we expected to alleviate concerns about inadvertent masking effects, illusory motion effects, and encoding disruption.

## 2. Method

### 2.1. Participants

The 32 participants (13 women, 19 men) were paid volunteers 19–48 years of age ($M = 26$, $SD = 7.4$). All signed a consent form approved by the MIT Committee on the Use of Humans as Experimental Subjects. Prior to experimentation, we assessed participants' comic reading experience using the "Visual Language Fluency Index" (VLFI) questionnaire, which asked participants to rate their frequency and expertise with reading comic books, comic strips, graphic novels, and Japanese comics, as well as drawing comics, both currently and while growing up (for more information, see Cohn et al., 2012). The "VLFI score" generated from this questionnaire has been shown to correlate with both behavioral and neurophysiological effects in the online comprehension of visual narratives (Cohn & Maher, 2015; Cohn et al., 2012). An "average" score along this metric falls near 12, with "low" being less than 7 and "high" above 20. Participants' fluency was average, with a mean score of 11.3 ($SD = 6.94$, range = 3.75–30.25). In addition, all participants rated their familiarity with *Peanuts* comics specifically as 1 (1 = low, 5 = high), though the stimuli used novel *Peanuts* sequences created for experimental purposes (described below).

Participants were replaced if their yes responses or coherence ratings were more than 2 *SD* above or below average. This was the case for two participants (one responded yes lower than acceptable, and one rated coherence almost exclusively with the highest rating).

### 2.2. Materials

Novel sequences were created by recombining black and white panels scanned from *The Complete Peanuts* volumes 1 through 6 (1950–1962) by Charles Schulz (Fantagraphics Books, 2004–2006). All text was edited out of the panels. *Peanuts* comics were chosen because (1) they have systematic panel sizes and content with repeated characters and situations; (2) their content is recognizable to most people; (3) there is a large corpus of sequences to draw from; (4) they feature fairly consistent and recurrent themes (various sports, building snowmen, Lucy skipping rope, etc.); and (5) they have been previously used in studies of VNG (e.g., Cohn et al., 2012). The present experiment used a total of 204 6-panel *Peanuts* strips.

The independently manipulated factors were presentation rate, first-panel duration, and panel ordering. Presentation rate was 500 or 1000 ms/panel. First-panel duration was either 500 or 1000 ms on 500 ms/panel trials and either 1000 or 2000 ms on 1000 ms/panel trials. Panel ordering was either ordered or mixed. Mixed trials switched the positions of two panels. Switching occurred in all six possible ways using panels in ordinal position 2 to ordinal position 5. A *switch distance* of one panel altered positions 2 & 3, 3 & 4, and 4 & 5. A distance of two panels used switches between positions 2 & 4 and 3 & 5, while reversal of positions 2 & 5 resulted in a switch distance of three panels.

Narrative constituent structures were coded for all sequences using theoretical diagnostic tests outlined by VNG (Cohn, 2013a, 2013b, 2014), which probe the coherency of constituents by moving, omitting, or substituting individual or groups of panels. Such tests are similar to those used for decades in linguistics to test the structure of syntax (e.g., Cheng & Corver, 2013). These tests were referenced against data from "segmentation tasks" (Gernsbacher, 1985) used in a prior study (Cohn et al., 2014) whereby participants were asked to draw lines between panels that would best divide the sequence into two parts (i.e., identify maximal constituent breaks). Across all sequences, many patterns of constituent structure were used. However, our primary analysis focused on sequences with two major constituents, as in Cohn et al. (2014), with the boundary between groups falling after the second, third, or fourth panel (34 sequences total). This ensured that not all switches between ordinal positions aligned with the same violations within versus between constituents. For example, the one-panel switch distance of altering positions 2 & 3 would fall within constituents if the boundary fell after the third panel, but would cross constituents if the boundary fell after the second panel. Thus, depending on the constituent structure of the sequence, both one and two panel switch distances could either occur within or between constituents, and all three panel switches (2 & 5) occurred between constituents.

### 2.3. Procedure

After obtaining informed consent, the participant was seated in a normally illuminated room in front of the computer and began testing. Each trial presented one 6-panel visual sequence one image at a time. The participant was asked to respond whether the panels were in order (with the 'y' key) or not (with the 'n' key), and then to rate the coherence of the panels as a narrative on a scale of 1–7, with 7 representing full comprehension.

There were 12 practice trials at 2000 ms/panel to begin the experiment, followed by 8 test blocks of 24 trials each. The test blocks alternated between 1000 and 500 ms/panel, starting with 1000 ms/panel. Participant number determined which block of images came first, e.g., participant 1 saw Block A first and participant 8 saw Block H first (followed by Blocks A, B, etc.). Each 24-trial block was pseudo-randomly arranged to include 12 ordered trials and 12 mixed trials, with no more than four trials of the same type in a row. Of the 12 mixed trials, there were two trials for each of the six switch distances. Six ordered and six mixed trials were randomly selected to present the first panel for twice as long as the following panels. The trials that had doubled first-panel durations for the first 16 participants had single first-panel durations for the second 16 participants, and vice versa. The trials that were ordered and mixed alternated every eight participants. This combination of counterbalancing and participant rotation resulted in each sequence being viewed under each order condition and each first-panel duration condition.

### 2.4. Apparatus

The experiment was programmed with MATLAB 8.3 and the Psychophysics Toolbox extension (Brainard, 1997), version 3, and was run on a Mac mini with 2.4-GHz, Intel Core 2 Duo processor. The Apple 17-in. CRT monitor was set to a 1024 × 768 resolution, with a 75-Hz refresh rate. The room was normally illuminated. Timing precision and stimulus presentation were controlled with the Stream package for MATLAB (Wyble, 2013).

### 2.5. Analyses

Repeated-measures analyses of variance (ANOVAs) were carried out on accuracy as a function of presentation rate and first-panel duration. Effect size was calculated using generalized eta squared (Bakeman, 2005).

The second response collected was coherence rating. This was judged on a 7-point Likert scale, with 1 representing no coherence and 7 representing full coherence and comprehension. The factors of trial type (mixed/ordered) and presentation rate were examined with respect to coherence ratings.

The relationships between VLFI score and both coherence rating and accuracy were analyzed to see if expertise contributed to visual narrative comprehension and detection of switched panels.

A fourth analysis was conducted on switch distance to determine how accuracy and coherence ratings were affected by panel switches that were adjacent compared to switches that were further apart. The factors of presentation rate, first-panel duration, and switch distance were used here. Switch distance and accuracy were factors in a second analysis of coherence to assess how accuracy and coherence interacted with respect to switch distance. Additionally, we examined how switches between and within constituents impacted accuracy and coherence ratings using switch distance and constituency (between/within) as factors.

## 3. Results

The four different groups of eight participants, containing different combinations of first-panel duration and ordering for each sequence,

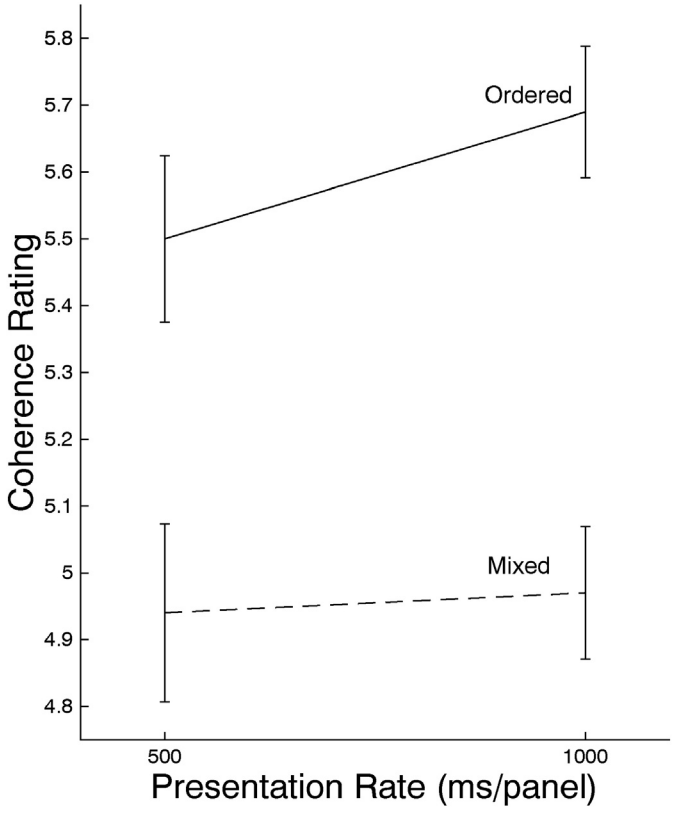


**Fig. 2.** Coherence rating as a function of panel presentation rate and ordering. Error bars are standard error of the mean.

did not produce different results from each other. In a two-way mixed ANOVA of participant group and presentation rate on accuracy, there was only a main effect of rate, $F(1,28) = 19.29$, $p < 0.001$, $\eta G^2 = 0.18$.

Examination of accuracy in a two-way ANOVA of presentation rate and first-panel duration revealed no effect of first-panel duration, but a strong effect of presentation rate, $F(1,31) = 9.02$, $p = 0.005$, $\eta G^2 = 0.077$. As expected, participants exhibited greater accuracy with longer exposure to the stimuli. A planned one-sample t-test comparing accuracy with chance (0.5) confirmed that participants were significantly above chance at both presentation rates, $ts(31) > 6.79$, $p < 0.001$.

Over the course of eight 24-trial blocks, participants showed no significant improvement between the first two and the last two blocks (quartiles). A Rate × Quartile ANOVA on accuracy only produced a significant main effect of rate, $F(1,31) = 9.02$, $p = 0.005$, $\eta G^2 = 0.07$.

### 3.1. Coherence ratings

Mean coherence ratings as a function of ordering and presentation rate can be seen in Fig. 2. Coherence ratings on a scale of 1 (no narrative comprehension) to 7 (full narrative comprehension) produced main effects of presentation rate, $F(1,31) = 4.6$, $p = 0.04$, $\eta G^2 = 0.05$, and trial type (ordered/mixed), $F(1,31) = 151.9$, $p = 0.001$, $\eta G^2 = 0.64$, with no interaction ($p = 0.11$). Both longer exposure time and correct panel ordering independently produced greater coherence. Coherence and accuracy did not significantly correlate across participants for either presentation rate, $rs < 0.11$, $ps > 0.51$.

### 3.2. Effect of expertise

Correlations between continuous VLFI scores and coherence ratings were not significant in the mixed and ordered conditions. Similarly, modeling linear regressions with a constant term did not produce significant effects of VLFI on coherence in the ordered, $F(1,30) = 2.75$, $p = 0.11$, $R^2 = 0.08$, and mixed conditions, $F(1,30) = 0.87$, $p = 0.35$, $R^2 = 0.03$. There was, however, a significant inverse correlation between continuous VLFI score and accuracy in the mixed condition, $r = -0.45$, $p = 0.01$, which may indicate that more experienced comics readers better tolerate panel switching. Corresponding with this significant correlation, linear regression with a constant term revealed a significant effect of continuous VLFI on accuracy in the mixed condition, $F(1,30) = 7.43$, $p = 0.01$, $R^2 = 0.199$, but not in the ordered condition, $F(1,30) = 0.3$, $p = 0.59$, $R^2 = 0.01$.

### 3.3. Switch distance

#### 3.3.1. Accuracy

A comparison between overall accuracy on the ordered trials and on subdivisions of the mixed trials assessed how far apart panels could be switched before narrative integration failed. When split into groups according to the distance panels were switched (immediately adjacent, 2-panel distance, or 3-panel distance) and analyzed in a Switch Distance × Presentation Rate × First-panel duration ANOVA, there was a strong effect of switch distance, $F(3,93) = 67.04$, $p < 0.001$, $\eta G^2 = 0.33$, and no effect of Presentation Rate or interaction. Immediately adjacent switches ($M = 0.49$, $SE = 0.03$) resulted in significantly worse order discrimination accuracy than both 2-panel ($M = 0.56$, $SE = 0.03$) and 3-panel switches ($M = 0.72$, $SE = 0.04$), $ts(31) > 4.21$. Correct detection of disorder with 3-panel switches was not significantly different from correct detection of ordered trials ($M = 0.79$ $SE = 0.02$), $p = 0.17$. First-panel duration did not impact these results.

There was an effect of panel switch distance on order discrimination accuracy compared to ordered trials, as shown in Fig. 3. As switch distance increased, participants were better at discriminating between ordered and mixed sequences. Accuracy measures were grouped based on the absolute difference between switched panels, but the temporal location of the switch impacted accuracy. Though accuracy was similar within each of the three switch distances, paired t-tests comparing accuracy between conditions within a switch distance group revealed that both the 2–3 ($p = 0.045$) and 3–4 conditions ($p = 0.014$) produced lower accuracy than the 4–5 condition. In the two-panel switch condition, there was no difference between 2 and 4 and 3–5, $p = 0.76$. Thus, one-panel switches that occurred early in a trial may have been harder to discern as unordered compared to later switches.

#### 3.3.2. Coherence

When the mixed condition was included in a two-way ANOVA examining Presentation Rate (2) and Switch Type (7, including ordered trials as 0-switch distance), There was a strong effect of switch type on coherence rating $F(6186) = 29.66$, $p < 0.001$, $\eta G^2 = 0.31$. Presentation rate did not have a significant effect on coherence rating in this analysis.

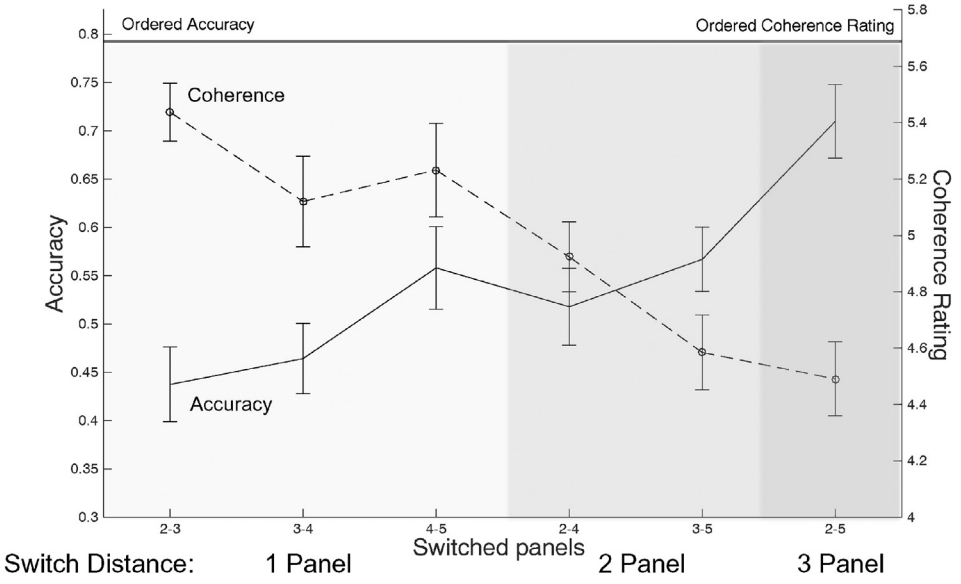


**Fig. 3.** Accuracy of discriminating mixed sequences (solid line) and coherence ratings (dashed line) of those sequences as a function of panel switch distance. Error bars are standard error of the mean. The solid horizontal reference line across the top of the figure shows both level of accuracy and coherence on ordered trials. Shaded regions depict different distances between the switched panels in the sequence.

Among the mixed trials, a Switch type × Accuracy ANOVA produced effects of both factors on coherence rating (Switch type: $F(5155) = 10.65$, $p < 0.001$, $\eta G^2 = 0.094$; Accuracy: $F(1,31) = 103.01$, $p < 0.001$, $\eta G^2 = 0.57$). When the narrative was ordered, participants responding "yes" were more likely to give a high coherence rating ($M = 6.08$). If they said "yes" to mixed trials, indicating they thought the sequence was correctly ordered, they also provided high coherence ratings ($M = 5.71$), compared to when they correctly responded "no," regardless of whether the sequence was ordered ($M = 3.97$) or mixed ($M = 4.21$). Coherence thus was a direct byproduct of perceived order.

*3.3.3. Constituent boundary effects*

Finally, we further analyzed our switched panels by examining a subset of our stimuli for switched panels that did or did not cross the boundary between constituents. Switches within constituents ($M = 0.43$, $SE = 0.029$) produced poorer discrimination performance than those between constituents ($M = 0.57$, $SE = 0.031$). In a Constituency (between/within) × Presentation Rate ANOVA of accuracy, we found significant effects of Constituency, $F(1,31) = 58.05$, $p = 0.001$, $\eta G^2 = 0.268$ and Presentation Rate, $F(1,31) = 11.369$, $p = 0.002$, $\eta G^2 = 0.111$.

Coherence ratings also differed based on switches across constituent boundaries, but in the opposite manner. Within-constituent switches ($M = 5.20$, $SE = 0.130$) produced higher ratings than between-constituent switches ($M = 4.78$, $SE = 0.117$). A Constituency × Duration ANOVA of coherence ratings produced an effect of Constituency, $F(1,31) = 34.26$, $p = 0.001$, $\eta G^2 = 0.206$.

Switch distances differentially impacted accuracy based on whether they occurred between or within constituents, as shown in Fig. 4. Namely, a greater switch distance facilitated order discrimination when the switched panels belonged to different constituents. A Constituency × Switch Distance (1 or 2 steps) ANOVA of accuracy revealed main effects of both Switch Distance, $F(1,31) = 9.88$, $p = 0.004$, $\eta G^2 = 0.109$, and Constituency, $F(1,31) = 12.67$, $p = 0.001$, $\eta G^2 = 0.118712$. There was also a significant interaction, $F(1,31) = 7.58$, $p = 0.009$, $\eta G^2 = 0.065$, which emerged because accuracy was much higher when the two switched panels were two panels apart between constituents ($M = 0.59$, $SE = 0.03$) compared to one panel apart between constituents ($M = 0.44$, $SE = 0.03$), two panels apart within constituents (M = 0.44, SE = 0.04), and one panel apart within constituents ($M = 0.41$, $SE = 0.03$). A two-way ANOVA of coherence confirmed main effects of Constituency, $F(1,31) = 5.77$, $p = 0.022$, $\eta G^2 = 0.06$, and Switch Distance, $F(1,31) = 33.69$, $p = 0.001$, $\eta G^2 = 0.29$.

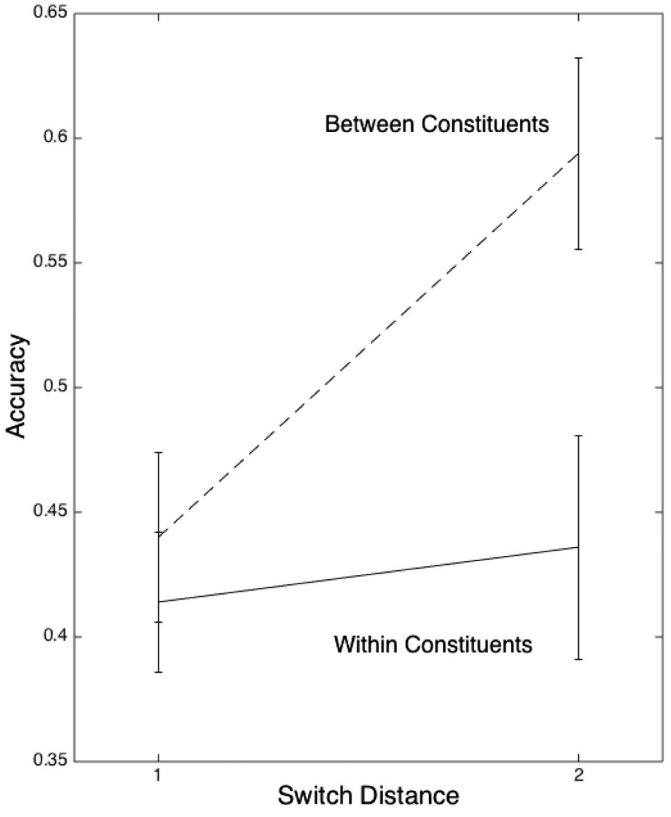


**Fig. 4.** Mean accuracy at determining a sequence was mixed when the two switched panels were 1 or 2 panels apart and within or between constituent structures. Error bars are standard error of the mean.

## 4. Discussion

Our main goal in the present experiment was to assess how observers integrated six sequential images into a narrative when two central images were switched in position either within or between a constituent structure. To accomplish this, we used two different rates of presentation (500 and 1000 ms/panel) with the same or double first-panel durations. We contrasted ordered trials, which presented the panels in their intended order, and mixed trials, which presented the sequence with two of the four interior panels switched. Overall, we found that viewers were able to distinguish ordered from mixed sequences with high proficiency, and their judgment of ordering directly impacted their rating of the narrative's coherence. Yet, both accuracy and coherence ratings were modulated by constituent structure, with switches between constituents more impactful than those within constituents. The rate of panel presentation affected their ability to discriminate order, with 1000 ms/panel allowing for nearly twice the sensitivity to order as 500 ms/panel, but the longer first-panel durations had no impact on accuracy. Finally, coherence of the sequence decreased as the distance between switched panels increased. Below, we discuss the implications of these findings in greater depth.

Longer panel durations allowed participants to more accurately judge whether the sequence was ordered or mixed. It was unsurprising that participants performed well when given 1000 ms to view each panel, since prior reported self-paced viewing times ranged from 700 to 2000 ms (Cohn & Paczynski, 2013; Cohn & Wittenberg, 2015). Since 500 ms/panel was below that range, we expected, and found, a decrease in sensitivity to ordering. Nevertheless, participants still detected order at above chance levels with the shorter viewing time. This suggests that readers of comics can assess the basic coherency of visual sequences at rapid paces. Previous RSVP tasks have generally used letters, numbers, words, or pictures as stimuli, and required participants to detect one or more targets in the presentation (e.g., Potter et al., 2014). Few studies have asked participants to globally integrate the presented stimuli into a narrative. One study presented words in scrambled or unscrambled sentences to assess the impact of sentence processing on the attentional blink (Potter, Nieuwenstein, & Strohminger, 2008). Meaningful sentence structure failed to promote detection of a second target word within the attentional blink. Whether the attentional blink would also affect detection of multiple target scenes in visual narratives is unknown, but worth comparative investigation.

While coherence was affected by presentation rate, modulation of a sequence's first-panel duration had no effect on the sequence. Previous work, including some using the same stimuli, has consistently found that the first unit of both verbal and visual narratives are viewed at longer durations than subsequent units (Cohn & Paczynski, 2013; Cohn, 2014; Foulsham et al., submitted for publication; Glanzer et al., 1984; Haberlandt, 1984) with this extra time devoted to "laying a foundation" for the discourse (Gernsbacher, 1990). We therefore presented the first panels at durations either the same or twice as long as the rest of the sequence, yet found this manipulation had no effect on order judgment or coherence. This suggests that, while viewers may prefer a longer duration to acquire initial information about a narrative when it is under their own control, inhibiting this process does not negatively impact understanding of the sequence's order. The relatively slow rates of

presentation may have allowed participants to use information in the later panels as readily as in the first panel, attenuating any possible impact. Testing faster rates of presentation could confirm this hypothesis.

For mixed sequences, greater switch distances resulted in generally better detection of order and worse ratings of coherence (Fig. 3). The adjacency effect was therefore strong in the present study, as previously observed by Cohn (2014) for drawn visual sequences. One possible explanation for this result is that, with a switch distance of only one panel, local cues were more likely left intact between adjacent panels, allowing for more effortless integration of the narrative's scenes, particularly when the switch came early in a sequence. Since the panel-switching was more salient, coherence was rated as low.

Examination of trials in which panel-switching caused constituent boundaries to be breached revealed that the ordinal switch distance alone did not motivate the discrimination and coherence of all rearranged panels. Switches between constituents were detected more accurately and rendered less coherent than those within constituents, consistent with findings of processing costs for the violation of narrative constituents (Cohn et al., 2014). These effects also interacted with the distance of the switched panels, which manifested when comparing one- and two-panel switch distances that could appear either within or between constituents. The switching of panels within and between constituents did not differ for adjacent panels, and was similar for two-panel switch distances within constituents. However, participants were significantly more accurate for two-panel switch distances that crossed between constituents and rated these as less coherent sequences. These results suggest enhanced detection of greater switch distances when they clearly violate constituent structures, above and beyond switches of ordinal position alone. Such findings provide further evidence for the segmentation of visual narratives into constituent structures (Cohn, 2013b; Gernsbacher, 1985), indicating that the preservation of constituent structure is likely critical to visual narrative grammar.

These results greatly expand on the work of Inui and Miyamoto (1981), who *only* switched adjacent panels, and found reasonably high accuracy rates (>62%) for panel durations with a *maximum* of 300 ms. Our results suggest that coherence degrades more substantially, and detection of ordering improves, if the violations extend beyond immediately adjacent panels, which we would expect to be compounded at faster durations. Such findings reinforce that sequential image comprehension must involve integration of panels across a global narrative context (Cohn, 2013b; Cohn et al., 2012) that extends beyond the linear relationships between images (Magliano & Zacks, 2011; McCloud, 1994).

Studies of event cognition have long shown that humans automatically impose hierarchic relationships onto several aspects of human cognition, whether they are event sequences (Jeon, 2014; Radvasnky & Zacks, 2014), language (Davis & Johnsrude, 2003; Patel, 2003), music (Koelsch, Rohrmeier, Torrecuso, & Jentschke, 2013; Patel, 2003), visual percepts (Bahlmann, Schubotz, Mueller, Koester, & Friederici, 2009; Tettamanti et al., 2009), or strictly conceptual, as in the case of mental arithmetic (Makuuchi, Bahlmann, & Friederici, 2012; Maruyama, Pallier, Jobert, Sigman, & Dehaene, 2012). This domain-general tendency to perceptually organize temporally extended stimuli requires segmentation of events. According to Event Segmentation Theory (Zacks, Speer, Swallow, Braver, & Reynolds, 2007) active ongoing event models are continually updated in relation to incoming sensory information across multiple timescales, allowing for perceptual prediction, a key facilitator of hierarchical temporal processing.

The hierarchic narrative structures explored here interface with event models, given that the function of a narrative structure is to package semantic understandings into a coherent order (Cohn, 2013b). It is worth noting that hierarchy in event structures is often determined via tasks in which comprehenders consciously segment events into coarse- and fine-grained divisions, which are sometimes correlated with neural activity to those same events (Zacks et al., 2001). While this method is effective for showing that hierarchical processing can be measured both in the brain and through behavior, it does not address the coherency of those units, or the costs of violating them. Our results suggest that, at least in the context of visual narratives, event models are flexible enough that alteration of the intended order of events within narrative constituents matters less than a change in events between constituents. Future research on event cognition should determine if the generality of hierarchical processing across domains extends to violations between and within constituent structure differentially affecting coherence and comprehension, such as film and music. Additionally, further exploration of the relationship between hierarchic structure in events and the narrative structure by which they are conveyed is needed to better quantify the determinants of coherence and understanding.

In sum, our present research demonstrates that readers of visual narratives can successfully integrate the global understanding of sequentially presented scenes into a coherent narrative when viewing panels for only 500 ms each and that violations of constituent structure negatively impact coherence. Future research should push beyond coherence and investigate the temporal limits of narrative comprehension using faster rates and longer sequences, in combination with manipulations targeting the narrative or semantic structure of sequences (e.g., Cohn et al., 2012), or internal aspects of panel content (e.g., Cohn & Paczynski, 2013; Cohn & Maher, 2015). Additionally, while we found no effect of first-panel duration on sensitivity to panel ordering, there are likely some semantic situations in which longer viewing times do enhance comprehension (Cohn, 2013b). Determining the conditions under which longer viewing times promote narrative comprehension would help us refine our understanding of visual narrative comprehension and how it is studied in relation to event patterns and sequences in other domains of cognition.

## Acknowledgments

This research was supported by a National Institutes of Health Grant MH47432. The authors report no conflict of interest involving this research. We thank Mary C. Potter for her comments on an earlier draft. Fantagraphics Books is thanked for their generous donation of *The Complete Peanuts*.

## References

Bahlmann, J., Schubotz, R. I., Mueller, J. L., Koester, D., & Friederici, A. D. (2009). Neural circuits of hierarchical visuo-spatial sequence processing. *Brain Research*, *1298*, 161–170 http://doi.org/10.1016/j.brainres.2009.08.017.

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior Research Methods*, *37*, 379–384.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*, 433–436.

Cheng, L., & Corver, N. (2013). *Diagnosing syntax*. Oxford University Press.

Clark, H. H. (1996). *Using language*. Cambridge, UK: Cambridge University Press.

Cohn, N. (2013a). *The visual language of comics: Introduction to the structure and cognition of sequential images*. London, U.K.: Bloomsbury.

Cohn, N. (2013b). Visual narrative structure. *Cognitive Science*, *37*(3), 413–452. http://dx.doi.org/10.1111/cogs.12016.

Cohn, N. (2014). You're a good structure, Charlie Brown: The distribution of narrative categories in comic strips. *Cognitive Science*. http://dx.doi.org/10.1111/cogs.12116.

Cohn, N., and Bender, P. Drawing the line between constituent structure and coherence relations in visual narratives. (submitted for publication)

Cohn, N., & Maher, S. (2015). The notion of the motion: The neurocognition of motion lines in visual narratives. *Brain Research*, *1601*, 73–84. http://dx.doi.org/10.1016/j.brainres.2015.01.018.

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

Cohn, N., & Wittenberg, E. (2015). Action starring narratives and events: Structure and inference in visual narrative comprehension. *Journal of Cognitive Psychology*, *27*(7), 812–828. http://dx.doi.org/10.1080/20445911.2015.1051535.

Cohn, N., Jackendoff, R., Holcomb, P. J., & Kuperberg, G. R. (2014). The grammar of visual narrative: Neural evidence for constituent structure in sequential image comprehension. *Neuropsychologia*, *64*, 63–70. http://dx.doi.org/10.1016/j.neuropsychologia.2014.09.018.

Cohn, N., Paczynski, M., Jackendoff, R., Holcomb, P. J., & Kuperberg, G. R. (2012). (Pea)nuts and bolts of visual narrative: Structure and meaning in sequential image comprehension. *Cognitive Psychology*, *65*, 1–38. http://dx.doi.org/10.1016/j.cogpsych.2012.01.003.

Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews. Neuroscience*, *3*(3), 201–215. http://dx.doi.org/10.1038/nrn755.

Davis, M. H., & Johnsrude, I. S. (2003). Hierarchical processing in spoken language comprehension. *Journal of Neuroscience*, *23*(8), 3423–3431 (http://doi.org/23/8/3423 [pii]).

Foulsham, T., Wybrow, D., & Cohn, N. Reading without words: Top-down attention in the viewing of comic strips. (submitted for publication)

Gernsbacher, M. A. (1985). Surface information loss in comprehension. *Cognitive Psychology*, *17*, 324–363.

Gernsbacher, M. A. (1990). *Language comprehension as structure building*. Hillsadale, NJ: Lawrence Erlbaum.

Glanzer, M., Fischer, B., & Dorfman, D. (1984). Short-term storage in reading. *Journal of Verbal Learning and Verbal Behavior*. http://dx.doi.org/10.1016/S0022-5371(84)90300-1.

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

Haberlandt, K. (1984). Components of sentence and word reading times. In D. E. Kieras, & M. A. Just (Eds.), *New methods in reading comprehension research* (pp. 219–251). Hillsdale, N.J.: Erlbaum.

Hagmann, C. E., & Cook, R. G. (2013). Active change detection by pigeons and humans. *Journal of Experimental Psychology. Animal Behavior Processes*, *39*(4), 383–389. http://dx.doi.org/10.1037/a0033313.

Hagoort, P. (2003). How the brain solves the binding problem for language: A neurocomputational model of syntactic processing. *NeuroImage*, *20*, S18–S29. http://dx.doi.org/10.1016/j.neuroimage.2003.09.013.

Hinds, J. (1976). *Aspects of Japanese discourse*. Tokyo: Kaitakusha Co., Ltd.

Inui, T., & Miyamoto, K. (1981). The time needed to judge the order of a meaningful string of pictures. *Journal of Experimental Psychology: Human Learning and Memory*, *7*, 393–396.

Jeon, H. -A. (2014). Hierarchical processing in the prefrontal cortex in a variety of cognitive domains. *Frontiers in Systems Neuroscience*, *8*, 223 http://doi.org/10.3389/fnsys.2014.00223.

Koelsch, S., Rohrmeier, M., Torrecuso, R., & Jentschke, S. (2013). Processing of hierarchical syntactic structure in music. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(38), 15443–15448 http://doi.org/10.1073/pnas.1300272110.

Kunzle, D. (1973). *The history of the comic strip (Vol. 1)*. Berkeley: University of California Press.

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology*, *62*, 621–647. http://dx.doi.org/10.1146/annurev.psych.093008.131123.

Magliano, J. P., & Zacks, J. M. (2011). The impact of continuity editing in narrative film on event segmentation. *Cognitive Science*, *35*, 1489–1517. http://dx.doi.org/10.1111/j.1551-6709.2011.01202.x.

Magliano, J. P., Dijkstra, K., & Zwaan, R. A. (1996). Generating predictive inferences while viewing a movie. *Discourse Processes*. http://dx.doi.org/10.1080/01638539609544973.

Makuuchi, M., Bahlmann, J., & Friederici, A. D. (2012). An approach to separating the levels of hierarchical structure building in language and mathematics. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *367*, 2033–2045. http://dx.doi.org/10.1098/rstb.2012.0095.

Mandler, J. M., & Johnson, N. (1977). Remembrance of things parsed: Story structure and recall. *Cognitive Psychology*, *9*, 111–151.

Maruyama, M., Pallier, C., Jobert, A., Sigman, M., & Dehaene, S. (2012). The cortical representation of simple mathematical expressions. *NeuroImage*, *61*, 1444–1460. http://dx.doi.org/10.1016/j.neuroimage.2012.04.020.

McCloud, S. (1994). *Understanding comics: The invisible art. understanding comics*. New York, N.Y.: Harper Collins.

Neville, H. J., Nicol, J. L., Barss, A., Forster, K. I., & Garrett, M. F. (1991). Syntactically based sentence processing classes: Evidence from event-related brain potentials. *Journal of Cognitive Neuroscience*, *3*(2), 151–165.

Patel, A. D. (2003). Language, music, syntax and the brain. *Nature Neuroscience*, *6*(7), 674–681 http://doi.org/10.1038/nn1082.

Potter, M. C. (1976). Short-term conceptual memory for pictures. *Journal of Experimental Psychology. Human Learning and Memory*, *2*(5), 509–522.

Potter, M. C., & Hagmann, C. E. (2014). Banana or fruit? Detection and recognition across categorical levels in RSVP. *Psychonomic Bulletin & Review*. http://dx.doi.org/10.3758/s13423-014-0692-4.

Potter, M. C., Nieuwenstein, M., & Strohminger, N. (2008). Whole report versus partial report in RSVP sentences. *Journal of Memory and Language*, *58*(4), 907–915. http://dx.doi.org/10.1016/j.jml.2007.12.002.

Potter, M. C., Wyble, B., Hagmann, C. E., & McCourt, E. (2014). Detecting meaning in RSVP at 13 ms per picture. *Attention, Perception & Psychophysics*.

Radvansky, G. A., & Zacks, J. M. (2014). *Event cognition*. Oxford, UK: Oxford University Press.

Rensink, R. A. (2004). Visual sensing without seeing. *Psychological Science*, *15*, 27–32.

Rumelhart, D. E. (1975). Notes on a schema for stories. In D. Bobrow, & A. Collins (Eds.), *Representation and understanding* (pp. 211–236). New York, NY: Academic Press.

Saraceni, M. (2001). Relatedness: Aspects of textual connectivity in comics. In J. Baetens (Ed.), *The graphic novel* (pp. 167–179). Leuven: Leuven University Press.

Shipley, T. F., & Zacks, J. M. (2008). *Understanding events: From perception to action*. New York, NY: Oxford University Press.

Spivey, M. J., & Dale, R. (2006). Continuous dynamics in real-time cognition. *Current Directions in Psychological Science*, *15*, 207–211.

Tettamanti, M., Rotondi, I., Perani, D., Scotti, G., Fazio, F., Cappa, S. F., & Moro, A. (2009). Syntax without language: Neurobiological evidence for cross-domain syntactic computations. *Cortex*, *45*(7), 825–838 http://doi.org/10.1016/j.cortex.2008.11.014.

West, W. C., & Holcomb, P. J. (2002). Event-related potentials during discourse-level semantic integration of complex pictures. *Cognitive Brain Research*, *13*, 363–375. http://dx.doi.org/10.1016/S0926-6410(01)00129-X.

Wright, A. A., Katz, J. S., Magnotti, J., Elmore, C. L., Babb, S., & Alwin, S. (2010). Testing pigeon memory in a change detection task. *Psychonomic Bulletin & Review*, *17*(2), 243–249. http://dx.doi.org/10.3758/PBR.17.2.243.

Wyble, B. (2013). *Stream program for matlab*.

Zacks, J. M. (2014). *Flicker: Your brain on movies*. Oxford, UK: Oxford University Press.

Zacks, J. M., Braver, T. S., Sheridan, M. A., Donaldson, D. I., Snyder, A. Z., Ollinger, J. M., ... Raichle, M. E. (2001). Human brain activity time-locked to perceptual event boundaries. *Nature Neuroscience*, *4*(6), 651–655.

Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., & Reynolds, J. R. (2007). Event perception: a mind-brain perspective. *Psychological Bulletin*, *133*, 273.

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, *123*, 162–185. http://dx.doi.org/10.1037/0033-2909.123.2.162.