



Cognitive Science 49 (2025) e70050

© 2025 The Author(s). *Cognitive Science* published by Wiley Periodicals LLC on behalf of Cognitive Science Society (CSS).

ISSN: 1551-6709 online

DOI: 10.1111/cogs.70050

A Computational Framework to Study Hierarchical Processing in Visual Narratives

Aditya Upadhyayula,^a Neil Cohn^b

^a*Department of Psychological & Brain Sciences, Washington University in St. Louis*

^b*Department of Communication & Cognition, Tilburg University*

Received 13 May 2023; received in revised form 30 January 2025; accepted 17 February 2025

Abstract

Theories of visual narrative comprehension have advocated for a hierarchical grammar-based comprehension mechanism, but only limited work has investigated this hierarchy. Here, we provide a computational framework inspired by computational psycholinguistics to address hierarchy in visual narratives. The predictions generated by this framework were compared against behavior data to draw inferences about the hierarchical properties of visual narratives. A segmentation task—where participants ranked all possible segmental boundaries—demonstrated that participants' preferences were predicted by visual narrative grammar. Three kinds of models using surprisal theory—an Earley parser, a hidden Markov model (HMM), and an n-gram model—were then used to generate segmentation preferences for the same task. Earley parser's preferences were based on a hierarchical grammar with recursion properties, while the HMM and the n-grams used a flattened grammar for visual narrative comprehension. Given the differences in the mechanics of these models, contrasting their predictions against behavior data could provide crucial insights into understanding the underlying mechanisms of visual narrative comprehension. By investigating grammatical systems outside of language, this research provides new directions to explore the generic makeup of the cognitive structure of mental representations.

Keywords: Visual narratives; Hierarchical grammar; Computational psycholinguistics; Probabilistic Earley parser; Hidden Markov models; Language models

Correspondence should be sent to Aditya Upadhyayula, Department of Psychological & Brain Sciences, Washington University, CB 1125, One Brookings Dr, St. Louis, MO 63130, USA. E-mail: aditya.usa8@gmail.com

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

1. Introduction

How does a story emerge from a sequence of images? One answer is that it uses a structure to organize the comprehended information, consistent with what distinguishes music from sounds, sentences from a string of letters, and events from a continuous experience. A large body of work in psycholinguistics has dealt with how linguistic information is organized into meaningful structured representations (Chomsky, 1956, 1957, 2013; Dehaene, Meyniel, Wacongne, Wang, & Pallier, 2015; Marcus, 2013; Sportiche, Koopman, & Stabler, 2013). Recently, work investigating other nonverbal cognitive domains has also demonstrated the presence of a cognitive structure, such as in the comprehension of events (Zacks, Speer, & Reynolds, 2009; Zacks & Tversky, 2001), music (Jackendoff, 2009; Jackendoff & Lerdahl, 2006; Lerdahl & Jackendoff, 1983), and narratives (Cohn, 2013b; Mandler & Johnson, 1977). Formal characterization of such cognitive structures is, therefore, crucial to studying how information is transformed into a meaningful representation. Investigating grammatical systems—that is, rule-based systems for arranging items into possible permissible combinations—is one solution that can be used to better understand the organizational principles of these cognitive structures.

Grammar is a key component of many spoken languages and has been studied as far back as Panini's work on the Sanskrit language. More recently, across the last half-century, grammatical systems have received a prominent focus in linguistics (Chomsky, 1957, 2009; Hauser, Chomsky, & Fitch, 2002) and cognitive psychology (Friederici, 2011; Hagoort & Indefrey, 2014; Sprouse & Schouse, 2020). Recent work has also advocated for the presence of grammatical structures in other domains. For example, music comprehension has been thought to have a syntax and a hierarchical grammar (Jackendoff, 2009; Lerdahl & Jackendoff, 1996; Pearce & Rohrmeier, 2018; Rohrmeier & Pearce, 2018). More recently, substantial evidence has shown that a narrative grammar works alongside semantic structures as we read visual narrative sequences, like those in comics or picture books (Cohn, 2013; Cohn, Jackendoff, Holcomb, & Kuperberg, 2014; Cohn, 2020). In this article, we investigate the hierarchical structure of the grammar used in visual narrative comprehension using a novel method that combines behavioral research with computational psycholinguistics tools.

Research on narratives has long identified that narratives follow a structure in which individual units are combined into groupings during comprehension (Mandler & Johnson, 1977). This structure differs from the semantic and event structures in that the narrative structure reflects the order in which events are presented—a crucial aspect of storytelling, while the former structures reflect the meaningful aspects of stories, that is, the characters and the events they undertake. This difference has also been noted elsewhere in the literature (Cuddon & Habib, 2013; Holquist, 2003; Pantaleo, 2004).

Theories of story comprehension maintain that the meaning arises from grouping coherent conceptual units using discourse relationships (Bransford & Johnson, 1972; Brown, Brown, Brown, Yule, & Gillian, 1983; Halliday, Hasan, & Hasan, 1985; Van Dijk, 1977) or scripts and schemas (Schank & Abelson, 1975, 2013). The emphasis in these theories is that situational changes such as shifts in spatial locations, intentionality, and causal changes between images represent boundaries between groupings of units (Magliano, Miller, & Zwaan, 2001;

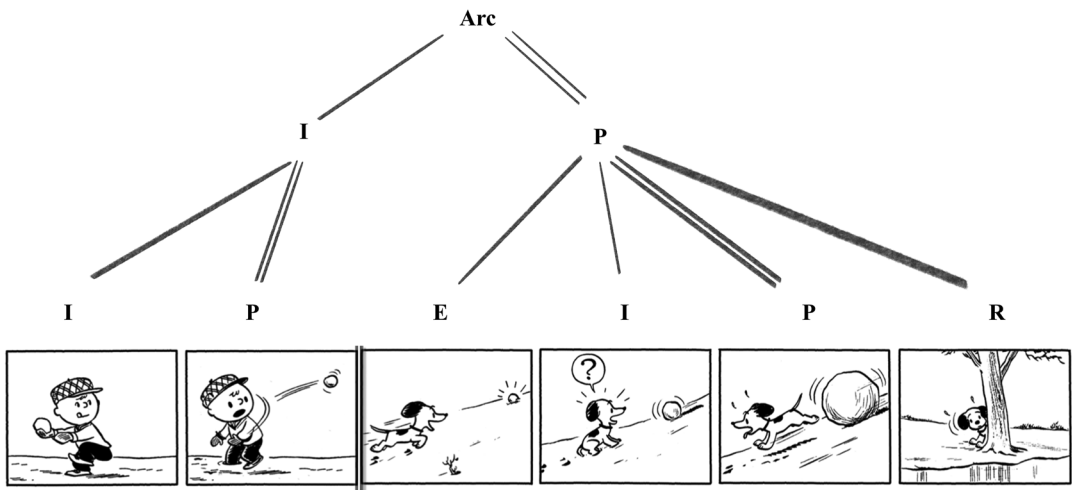


Fig. 1. Hierarchical visual narrative grammar in comics.

Magliano & Zacks, 2011; Zacks et al., 2009; Zwaan, Langston, & Graesser, 1995; Zwaan & Radvansky, 1998). The greater semantic discontinuity between the units often increases the likelihood of a boundary being present, which in turn reflects the grouping structure in comprehension. Nevertheless, such situational changes only capture pairwise coherence relationships between the contents of discourse units. As a result, theories focusing only on semantics are thus limited in accounting for several other properties of narrative comprehension such as the narrative roles played by units in relation to each other (Cohn, 2014), their global structure which might result in structural ambiguities (Cohn, 2013b), and cross-cultural differences of diverse narrative patterns (Cohn, 2023; Cohn & Kutas, 2017) among other factors.

To account for these properties, Cohn (2013a) proposes that a narrative structure with grammatical properties interfaces in parallel with the type of semantic processing described in other discourse theories. visual narrative grammar (VNG) posits that each image within a visual sequence plays a grammatical role in the narrative arc, typically consisting of an Establisher (E), Initial (I), Peak (P), and Release (R). An Establisher (E) introduces the relationship between the characters without acting on it. Initials (I) then depict the set-up, often through preparatory actions involving the characters. The Peak (P) depicts the climax of the constituent with the maximal and most informative part of the sequence. Finally, a Release (R) provides the aftermath or resolution of the climax and usually signals the end of the narrative. Sometimes, information in one panel is also prolonged into the subsequent panels. This is depicted by a Prolongation (L). A canonical schema of the narrative arc presents these categories in this E-I-P-R order (Cohn, 2013b, 2020). However, this narrative schema can be recursively combined to give rise to various combinations of narrative category orders and an emergent hierarchical structure.

Consider Fig. 1, which is a modified strip from “Peanuts” by Charles Schultz involving two principal characters Charlie Brown and his dog Snoopy. In this strip, Charlie Brown throws a

snowball, and his dog Snoopy chases the snowball. The ball then grows while rolling down the hill frightening Snoopy, who then hides behind a tree. The first panel is an Initial (I), motivated by the cues of the preparatory action of Charlie Brown reaching back the ball. The second panel is a Peak (P), where the throwing action manifests. The third panel presents a new relationship between Snoopy and the snowball, as an Establisher (E). Next, Snoopy realizes the mounting threat in another Initial (I), before doing the climactic Peak action (P) of running away. In the final panel, Snoopy hides behind the tree, in the aftermath of the scene, as the Release (R).

The surface sequence of this structure is I-P-E-I-P-R, which extends beyond (and would violate) the expectations of a canonical narrative schema. To allow for this, a key aspect of VNG is that in addition to canonically functioning as roles played by units (such as a panel in a comic), they also recursively characterize the relative roles played by whole groupings of panels (i.e., narrative *constituents*). For example, the group of the first two panels (Initial-Peak) forms an Initial (I) at a higher level of structure. This constituent is motivated by the “head” Peak panel within that constituent, that is, without Charlie Brown throwing the ball, the next four panels would not have happened. The remaining four panels then function as another constituent that assumes the role of a Peak (P) at the higher-level structure. These are again motivated by the subordinate Peak, which is the primary climax of the whole sequence. The result, therefore, is a hierarchical recursive structure, as in Fig. 1 (see Cohn, 2015 for a detailed example).

Through this hierarchical structure, VNG can account for structural properties like distance dependencies and structural ambiguities in visual narrative sequences—thus going beyond what is conveyed by pairwise meaningful associations. VNG follows a constructional approach similar to contemporary linguistic models, in contrast to the earlier grammatical approaches to stories which largely followed Chomskyan grammatical models (Mandler & Johnson, 1977; Rumelhart, 1975). For example, unlike Chomskyan models which posit procedural rules, VNG argues that narrative structures are encoded in long-term memory as declarative constructional schemas. In addition, unlike the ambiguous relationship between structure and meaning in prior story grammars (Mandler & Johnson, 1977), VNG posits that narrative structure explicitly operates in parallel to the semantic structures described in discourse theories (Loschky, Hutson, Smith, Smith, & Magliano, 2018, 2020; Zwaan & Radvansky, 1998). The narrative grammar provides a structural packaging for the meaningful information used in the mental model construction (Cohn, 2020).

Both behavioral and neurocognitive research have supported this parallel role of the narrative structure to semantic processing. For example, neural responses corresponding to semantic processing (the N400) were insensitive to the presence of narrative structure in the absence of coherent meaning (Cohn, Paczynski, Jackendoff, Holcomb, & Kuperberg, 2012). In reverse, neural responses sensitive to narrative patterns (the Lateral Anterior Negativity) were insensitive to manipulations of semantic incongruity (Cohn & Kutas, 2017).

Prior work studying the narrative structure has shown that narratives are grouped into constituents. For example, Cohn et al. (2014) introduced disruptions either within or between narrative constituents in visual sequences, while measuring participants’ neural responses using Event Related Potentials (ERPs). This technique followed classic “click” studies from

psycholinguistics used to study the constituents of sentences (Fodor et al., 1974; Fodor & Bever, 1965). These disruptions evoked neural responses consistent with syntactic processing in language. Moreover, differences arose between the neural responses to disruptions at and before the constituent boundaries, preceding any type of semantic discontinuity considered by discourse theories to trigger a constituent break.

Further work has also shown that narrative structure goes beyond semantic discontinuity at constituent boundaries. Cohn and Bender (2017) used a segmentation task where participants were asked to divide a visual narrative sequence into two parts and then continued segmenting all possible boundaries between panels. Similar discourse tasks are used in discourse studies (Mura, Petersen, Huff, & Ghose, 2013; Newton, 1973; Newton & Engquist, 1976) under the assumption that segmental boundaries are motivated by situational discontinuity (Zacks et al., 2009; Zwaan et al., 1995; Zwaan & Radvansky, 1998). Though semantic discontinuity did predict participants' first preferred segmental boundaries, such semantic changes were not as predictive as the hypothesized constituent breaks as determined by narrative category pairings. For example, bigrams such as R-E, P-I, R-I, and P-E violate the rules of VNG and were more predictive of segmental boundaries than situational discontinuity (Cohn & Bender, 2017).

The research on visual narrative comprehension has established the existence of narrative constituents, and that the narrative grammar informs how these constituent boundaries are perceived better than situational discontinuity alone. VNG posits that these boundaries naturally arise via hierarchical relationships between substructures that are part of a higher-level cognitive structure, as in Fig. 1. However, it is also possible that constituent structures could arise in a flattened structure, where a boundary could arise from adjacent units that do not typically co-occur. Indeed, Cohn and Bender (2017) found that the greatest predictor of segmentation in a visual sequence was the relative positioning of narrative categories (e.g., the most predictive cue for segmentation was an Establisher as the first category of a new constituent). Prior works are yet to confirm a distinction between a hierarchical grammar, which allows for recursively combining multiple elements, versus a flattened grammar, which lacks such recursion.

To address this issue, we introduce a computational framework for visual narrative comprehension inspired by research in computational psycholinguistics (Hale, 2001, 2016; Levy, 2008; N. J. Smith & Levy, 2013). This computational framework provides a proof of concept for studying the mechanistic processes underlying visual narrative comprehension. The computational models implemented here make use of a hierarchical grammar as described by the VNG to make specific predictions about visual narrative structure in a segmentation task. To our knowledge, few computational approaches have examined visual narrative processing (Baldassano et al., 2017, 2018; Franklin, Norman, Ranganath, Zacks, & Gershman, 2020; Gershman, Radulescu, Norman, & Niv, 2014; Laubrock & Dunst, 2020; Martens, Cardona-Rivera, & Cohn, 2020; Reynolds, Zacks, & Braver, 2007). Much of the work here focused on data-driven approaches to explain or generate events and narrative comprehension. In contrast, our approach uses the theoretical underpinnings of VNG combined with recent advances in computational psycholinguistics to explore a probabilistic Earley parser (Hale, 2001; Stolcke, 1994), a hidden Markov model (HMM) (Juang & Rabiner, 1991; Rabiner & Juang,

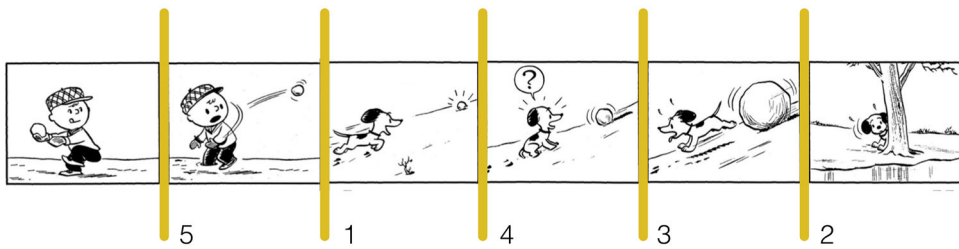


Fig. 2. Offline segmentation task. Participants were asked to rank all possible boundaries in their preferred order. The numbers indicate an example preferred order of segmentation.

1986), and an n-gram model as models of visual narrative comprehension. These models compute the probability of predicting the next unit in a sequence (typically words in a sentence) given the observed context. However, they differ in at least one aspect: Earley Parser makes use of a hierarchical grammar, while HMMs and n-grams use the linear (i.e., pairwise) transition probabilities between the observed states. Given this difference, these models can provide insight into the investigation of hierarchy in the narrative structure of visual sequences.

In the experiments to follow, we compare the performance of the probabilistic Earley parser, the HMM, and n-grams to previously collected behavioral data in a segmentation task to make specific predictions about the underlying narrative structure. Experiment 1 aimed to reanalyze the original findings of Cohn and Bender (2017), while Experiment 2 simulates the models' performances and compares against this behavioral data.

2. Experiment 1: Segmentation of visual narratives

Prior research in discourse theory has used both online and offline variations of the segmentation task to study discourse structure (Mura et al., 2013; Newton, 1973; Newton & Engquist, 1976). In the online version of the task, participants are asked to actively segment visual narratives as they are unfolding temporally (Huff, Meitz, & Papenmeier, 2014; Magliano et al., 2001; Magliano & Zacks, 2011). In contrast, the offline version of the task involves segmentation of narrative after presenting the whole sequence or movie clip (Cohn & Bender, 2017; Gernsbacher, 1985; Kurby & Zacks, 2012). Furthermore, offline segmentation tasks allow participants to rank their segmentation preferences, which could then be used to make inferences about the underlying narrative structure. In this study, we used the offline segmentation data from Cohn and Bender (2017), which had participants rank all possible boundaries in a comic strip in their preferred order (see Fig. 2).

Cohn and Bender (2017) analyzed the first two ranked boundary preferences to show that a boundary between two panels is predicted by their respective narrative category bigrams more than the semantic changes in each comic strip. We aim to analyze these data using a slightly different approach. We utilize the rank order preferences to compute a new metric called "boundary agreement"—a score informing how well participants agree that a given panel pair contains a first preferred boundary. According to VNG, panel pairs containing

noncanonical category orders such as R-E, R-I, P-E, and P-I are “illegal” and should not constitute a single constituent (because the canonical arc follows an E-I-P-R order), and thus these category pairings should represent the location of a boundary between constituents. Thus, if segmentation is influenced by narrative grammar, we should expect noncanonical categorical panel pairs to result in a higher boundary agreement compared to the canonical pairs.

3. Methods

3.1. *Stimuli and apparatus*

The data and code for this article are compliant with the open science transparency guidelines and are available at <https://osf.io/s2h5x/>. One hundred and twenty coherent black and white graphic sequences constructed from *The Complete Peanuts* volumes 1 through 9 were used in our study. We used panels without text to eliminate the undue influence of the written language. Each sequence comprised six panels that were combined from the individual panel database to create a coherent narrative. The generated sequences have also been used in various prior studies (Cohn et al., 2012; Cohn & Paczynski, 2013).

3.1.1. *Coding of narrative structure*

Each of the sequences was collaboratively coded by the authors following the rules and diagnostic tests of VNG (see Cohn, 2015 for examples). All the generated narrative structures comprised two constituents, with some panels belonging to the first and the remaining panels out of the total six belonging to the second constituent, respectively. Furthermore, these narrative structures were also validated by comparing them to an earlier empirical study where participants were asked to make a single segmentation into constituent sequences (Cohn et al., 2014). Overall, this resulted in 40 sequences of 2–4, 3–3, and 2–2 patterns each, with the constituent boundary appearing after the second, third, and fourth panels, respectively.

3.1.2. *Coding semantic coherence changes*

Prior research studying the semantic organization of narratives has demonstrated various salient dimensions along which participants have segmented visual sequences (see Zwaan and Radvansky, 1998 event indexing model for more details). Changes in spatial location, causality, and character changes between panels have been shown to significantly influence narrative segmentation. All 120 sequences were consequentially coded for the semantic changes from the spatial, causal, and character change perspectives. A change in characters between two panels was coded as 1 for a complete change in characters represented by the panels, 0.5 for a partial change, and 0 for no change. A similar coding scheme was applied for both spatial and causal changes. Thus, each panel was coded as a three-dimensional vector that represented how it differed from the previous panel along space, character, and causal dimensions.

3.2. Participants

A total of 54 comic readers (27 male, 27 female, mean age: 23) from the Tufts University undergraduate population carried out the segmentation task. All participants gave informed written consent according to Tufts University's Human Subjects Review Board guidance. Participants' comics reading experience was assessed using the "Visual Language Fluency Index" score that generated a fluency score based on a pretest questionnaire concerning their exposure to reading comics and other cultural factors (see Cohn, 2020 for more details). A high fluency score implies that the participant is proficient with reading comics, and vice-versa. On average, participants had a wide range of fluency scores (low = 4.38, high = 35.38), with an average fluency of 14.35 ($SD = 6.24$), which is average along an idealized range (<8 = low, $12-20$ = average, >20 = high).

3.3. Design and procedure

Participants were provided with a packet of the printed six-panel comic strips, with one full strip on each page, and a pencil. For each comic strip, they were instructed to draw a line between the panels which they thought would most intuitively divide the sequence into two parts and to label it with a "1." Each strip had six panels, and thus five possible boundary locations that divided the sequence. They were then instructed to continue drawing lines with labels to divide these segments, until all the five possible gutters were divided. As each division was numbered, it assigned a rank order to the segmentations, with a rank of 1 implying the highest preference for a constituent boundary, and 5 implying the opposite.

4. Data analysis

Boundaries with a consistently high rank imply a high agreement among participants for that boundary location. For each participant, the boundary preference was coded using the rank order. The first preferred boundary was given a rank of 1, and the final preferred boundary was given a rank of 5. For each of the five boundaries between panels in the 120 strips, we then used the rankings given by all 54 participants and computed a boundary agreement score. This score reflects the proportion of the time participants agreed that a first-preferred boundary exists between any two panels. For example, if the second boundary, that is, the boundary between the second and the third panel in a given comic strip had 30 out of 54 participants rating it as their first segmentation, the agreement score would be $30/54 = 0.55$. Similarly, if a given boundary has 10 out of 55 participants rating it as their first segmentation, the agreement score would be $10/54 = 0.18$. Thus, each possible boundary between the panel pairs in each strip had a continuous boundary agreement score ranging between 0 and 1. Additionally, this score facilitates comparison between human and computationally derived segmentation preferences which we describe later in Experiment 2. We predicted that panel pairs with noncanonical narrative category pairs (e.g., R-E, R-I, P-E, and P-I) would reflect a high boundary score compared to the canonical pairs.

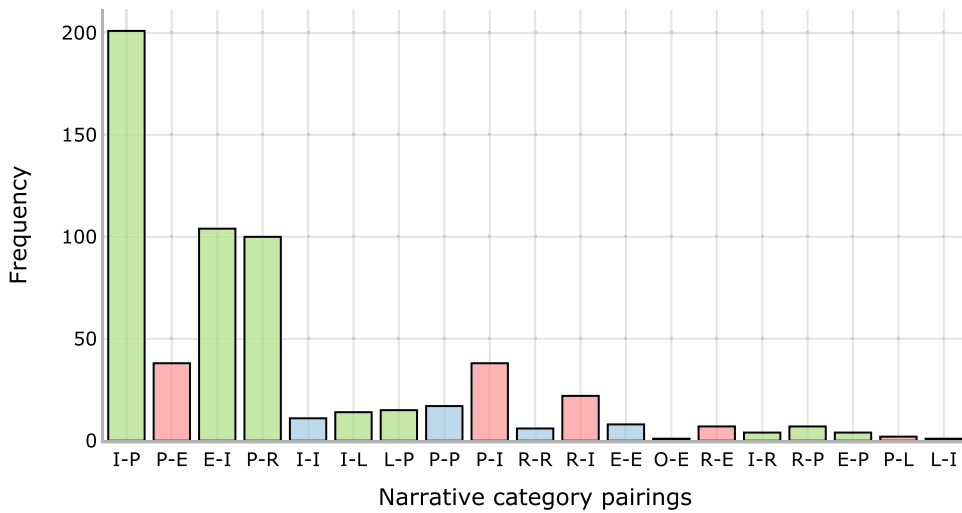


Fig. 3. Narrative category linear pairings across the 120 comic strips. Green-colored items represent canonical bigram pairs. Blue-colored items represent canonical repeat bigram pairs. The red-colored items represent illegal bigram pairs as per VNG.

Each sequence had six panels, and thus five possible boundary locations, five narrative category pairs, and five situational changes. See Fig. 3 for a histogram plot of the distribution of the narrative category pairs. Boundary agreement scores were computed at all five locations. To quantify the relationship between boundary agreement and the category pairs, these pairs were binned into three categories: illegal, legal (canonical), and legal (repeat). The illegal category consisted of pairs that cannot coexist within a single constituent as per the rules of VNG. These include pairs such as R-E, R-I, P-E, and P-I. The legal (repeat) category consisted of pairs that are repetitive and are allowed as “conjunctions” per VNG (Cohn, 2013a). These included pairs such as E-E, I-I, P-P, and R-R. The remaining pairs were binned as legal (canonical).

5. Results

We found that participants preferred the theoretically identified constituent breaks as their first boundary preference. As in Fig. 4a, participants ranked the boundary between the constituents as their first boundary about 51.5% of the time compared to boundaries elsewhere within constituents (about 12.35% of the time). Furthermore, the boundary agreement on average for the theoretically derived constituent boundary, that is, between constituents ($M = 0.56$, $SD = 0.11$) significantly differed compared to within constituents ($M = 0.18$, $SD = 0.08$) in the comic strip ($t(598) = 18.10$, $p < .001$)—see Fig. 4b.

We also analyzed how the boundary agreement score is predicted by semantic changes in the comic panels (e.g., location, character, and causal changes across the panels), and the

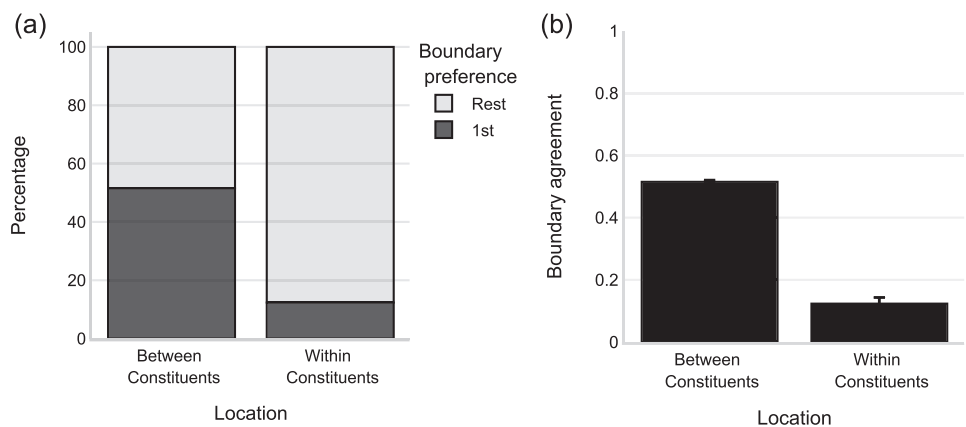


Fig. 4. (a) Boundary preference as a function of location. Light gray regions show the proportion of time participants did not rate the boundary as their first preference. (b) Boundary agreement as a function of boundary location. Error bars show the standard error of the mean.

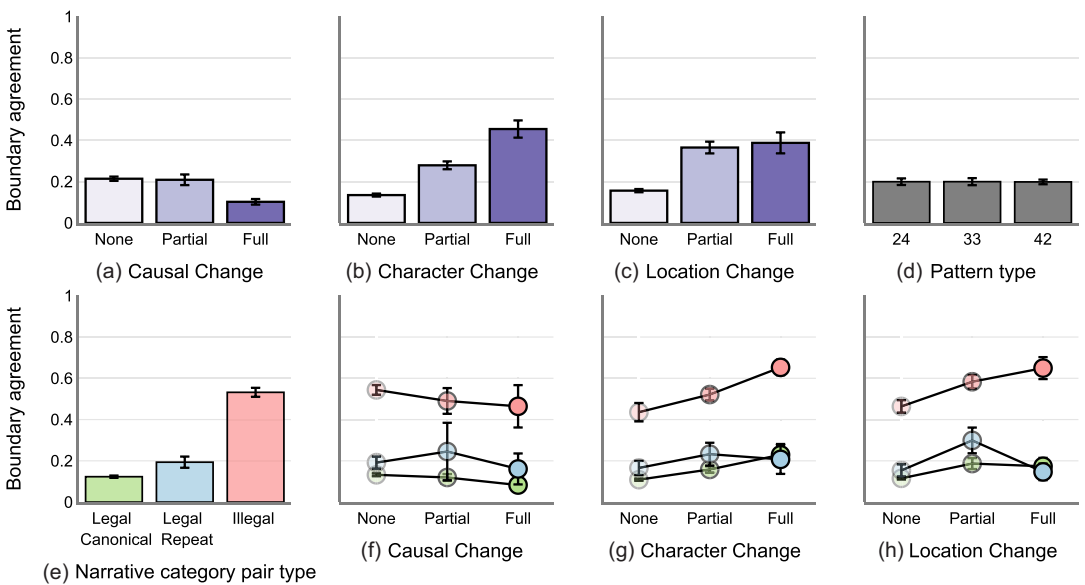


Fig. 5. Boundary agreement as a function of semantic, narrative category pairs, and pattern-type changes. Color-coded data points in panels f, g, and h reflect the syntax pair type plotted in panel e. Situational changes in these panels had higher boundary agreement for illegal bigram pairs compared to the other conditions. Error bars and shaded regions represent standard error.

type of narrative category bigram pairs (e.g., illegal, legal [canonical], and legal [repeat]), as shown in Fig. 5. On average, illegal pairs resulted in higher boundary agreement compared to the canonical pairs. Boundary agreement increased with an increase in the degree of location

Table 1
Linear mixed-effects model estimates

Predictors	Boundary agreement			Random effects, <i>SD</i> <i>Boundary number</i>
	Estimates	<i>CI</i>	<i>p</i>	
Intercept	0.11	[0.04, 0.17]	<.001	0.141
Narrative category pair (Repeat)	0.02	[−0.02,0.06]	.408	—
Narrative category pair (Illegal)	0.29	[0.26, 0.33]	<.001	—
Cause (Partial)	−0.01	[−0.04,0.03]	.740	—
Cause (Full)	−0.03	[−0.06,0.00]	.085	—
Location (Partial)	0.06	[0.02, 0.09]	.001	—
Location (Full)	0.08	[0.01, 0.14]	<.001	—
Character (Partial)	0.05	[0.03, 0.08]	<.001	—
Character (Full)	0.12	[0.07, 0.16]	<.001	—
Constituent Pattern (3-3)	0.00	[−0.02,0.03]	.797	—
Constituent Pattern (4-2)	0.02	[−0.01,0.04]	.150	—
Marginal <i>R</i> ² /Conditional <i>R</i> ² : .507/.631, <i>N</i> _{obs} = 591				

Note. Significant effects are highlighted in bold.

and character changes resulted in a higher boundary agreement score. On the other hand, an opposite effect is observed for causal change between panels.

We fit a linear mixed model (estimated using REML from the “lmerTest” package in R) to predict boundary agreement. Narrative category pairs, Causal, Location, Character-specific changes, and Pattern type were input as the fixed effects. Interaction terms were excluded from the model to control for the variance inflation factors. Further, the five possible boundary locations in each of the six-panel strips were included as random effects at the intercept level. These were enumerated from 1 through 5 with 1 corresponding to the boundary following the first panel, and so on. No random slopes were included to facilitate the model convergence.

Overall, this model captured about 63% of the total variance (conditional *R*² = .631). Model estimates are reported in Table 1, see also Fig. S4 for the model estimate plots. Overall, changes in location, character, and the type of syntax pair were significant predictors of the boundary agreement. An ANOVA of the model estimates confirmed significant main effects of the narrative category pair type (*F*(2,578) = 165.57, *p* < .001), Location changes (*F*(2, 576) = 9.64, *p* = .001), and Character changes (*F*(2, 576) = 16.24, *p* < .001). There was no significant main effect of Causal change *F*(2, 576) = 1.48, *p* = .22). We discuss the implications of the lack of an effect of causal change on segmentation agreement in the General Discussion. Finally, there was no significant main effect of the pattern type in predicting boundary agreement scores (*F*(2, 576) = 1.14, *p* = .31).

6. Discussion

This experiment investigated whether participants’ segmentation of visual sequences is influenced by situational changes and/or a narrative structure. On average, participants rated

the expected boundary between two constituents as their top choice compared to the boundaries within a constituent, as in Fig. 4. Note that in this context, a 50% agreement is much higher than the chance level of 20% given that participants had five choices to choose from. A boundary agreement score was then computed to quantify how well participants agreed on a panel pair containing a preferred boundary.

Our results demonstrate that the boundary agreement score is predicted by the narrative category pairs. Particularly, the illegal pairs as characterized by VNG had higher boundary agreement compared to the other types of pairings within the narrative grammar (Fig. 5d). It is worth noting that despite being less frequent in the corpus (Fig. 3), illegal pairs were still significant predictors of the boundary agreement score. Furthermore, semantic changes—that is, changes in the scene locations and characters—also predicted boundary agreement (see Figs. 5a–c). Interestingly, there was no effect of causal changes in predicting the boundary agreement. We discuss the implications in the General Discussion. Overall, these results support the main findings of Cohn and Bender (2017) that narrative categories (e.g., E, I, P, R) were significant predictors of participants' segmentation preferences.

It is worth noting that, unlike the situational changes which are visible to a participant, the illegal category pairings such as R-I, R-E, P-I, and P-E are theoretical VNG constructs, which participants in this study were unaware of when segmenting the comic strips. The boundary agreement scores were higher at these illegal narrative category pairs even after controlling for semantic changes (Fig. 5). Therefore, a high boundary agreement among participants at such locations implicates a crucial role of narrative categories and constituent structures in visual narrative comprehension. However, the results of this segmentation task only implicate boundaries between panel pairs and leave open whether the constituents result from having a hierarchical structure as opposed to a flat organization of the narrative constituents. In the following experiment, we explored whether computational models of visual narrative comprehension could be used to study the associated hierarchical properties.

7. Experiment 2: Computational psycholinguistics in visual narratives

In computational psycholinguistics, the prediction of upcoming words in a sentence is characterized by surprisal theory (Hale, 2001, 2016; Levy, 2008; N. J. Smith & Levy, 2013), which models word prediction as a continuous probabilistic variable ranging from 0 (surprising) to 1 (near certainty). Surprisal of a word is proportional to the negative log probabilities for observing the word w_i in a given context and the previous words (Levy, 2008)—see Eq. 1. The context here refers to the sentence structure—which could reflect the organizational structure of the words. This definition implies that the surprisal is minimized to 0 if a given word is most certain to appear in a given context. Similarly, surprisal is maximized to infinity (∞) if a word is not to appear in a given context. Computational models of surprisal theory have been used to study syntactic structure in language comprehension (Hale, 2001; Levy, 2008; Linzen, Dupoux, & Goldberg, 2016; N. J. Smith & Levy, 2013).

$$s(w_i) \propto -\log(p(w_i | w_{\{1,2,\dots,i-1\}}, context)) \quad (1)$$

Because these models are not necessarily bound to words and language specifically, we can apply them to other modalities such as visual narrative sequencing. A few other applications include event composition (Tran, Collier, Le, Phi, & Pham, 2013), language modeling (Hale, 2001), human activity prediction (Qi, Jia, Huang, Wei, & Zhu, 2020), and speech recognition (Jurafsky et al., 1995). If segmentation preferences in visual narratives are influenced by the underlying narrative structures and schemas, they should be reflected in the computed surprisals that are informed by the same structures. We, therefore, investigated the segmentation data from Experiment 1 using a surprisal theory perspective to obtain boundary agreements computationally. In the context of visual narrative comprehension, surprisal for a given panel is proportional to the negative log probability of observing the panel given the previous panels and the context—analogue to the psycholinguistic surprisal definition. A panel with a high surprisal score implies a low probability of co-occurring with the previous panel in the sequence and thus suggests a constituent break. Therefore, a high surprisal score could likely inform an underlying constituent boundary. We predict that surprisal scores would be higher at panels following theoretically derived constituent boundaries compared to the panels preceding boundaries in the visual sequence.

Theories of visual narrative comprehension posit that both semantic and narrative grammar structures operate independently in parallel to create a coherent mental model (Cohn, 2020; Loschky et al., 2018, 2020; Zwaan & Radvansky, 1998). We incorporated both the semantic change information and the VNG to compute the surprisal values. Semantic changes included changes in spatial location, causality, and character changes at the panel level as described in Experiment 1, while the narrative syntax comprised the Establishers, Initials, Peaks, and Releases and their associated grammar. Hierarchical syntax trees along with semantic changes were implemented using a probabilistic Earley parser (Hale, 2001; Stolcke, 1994), while the narrative category pairwise pairings along with semantic changes were implemented using HMMs (Juang & Rabiner, 1991; Rabiner & Juang, 1986). A key aspect of both the Earley parser and the HMMs is that they can be used to model the semantic and syntactic information in parallel as proposed by the narrative comprehension theories. However, they operated on fundamentally different narrative structures. Earley parser used top-down information from the hierarchical syntax trees alongside the semantic change information to create a grammar for computing surprisals in the comic strips. HMMs were agnostic to the hierarchical structure and relied only on the pairwise narrative category pairings alongside the semantic changes to compute surprisals. Additionally, we also used n-grams in our analysis. These n-gram models only had access to either semantic information or syntax information, but not both unlike the Earley parser and the HMM. The surprisal estimates generated by the three models can be used to study hierarchical properties of visual narrative comprehension.

8. Methods

8.1. Earley parsers

Earley parsers (Hale, 2001; Stolcke, 1994) are top-down parsers that make use of context-free grammars to make predictions about the upcoming data in the sequence. At any given

moment, the state of the parser can be completely defined by the collection of the states that define the tree set. A state is a record of the current input string position, a grammar rule, a pointer indicating how much of the grammar rule has already been recognized, and the leftmost edge of the substring the rule generates. Broadly speaking, the Earley parser has three main stages: (1) predict, (2) scan, and (3) complete. In the predict stage, the parser adds new states from the possible set of rules in that position—this stage predicts the next unit in the sequence using these rules. The scan stage then looks for the next incoming unit in the input string and matches it against the state generated by the rules in the predict step. Finally, the complete step updates the set of rules that are already recognized. This is done iteratively from the start to the end of the sequence for every unit within the sequence.

The grammar rules describe how narrative categories are combined into constituents and organized into a hierarchical structure. As described earlier, a narrative Arc could consist of an Establisher (E), an Initial (I), a Peak (P), and a Release (R). Although illegal according to the VNG, pairings such as P-E, R-I, P-I, and R-E are often seen together in a sequence—see Fig. 1 for reference. VNG facilitates such occurrences by recursively combining each of these syntactical elements to form a higher-level constituent as shown in Fig. 1. Here, an Initial could be broken down into another Initial (I) and a Peak (P)—which is a perfectly acceptable rule as per VNG. Similarly, the second head-level peak can be broken down into E, I, P, and R—which is also acceptable as per VNG. This way, illegal pairs such as P-E at the second and the third panels can co-occur by belonging to different constituents through a hierarchical organization as seen in Fig. 1. Any unusual pairings at different levels of hierarchy could lead to higher surprisals. As an example, higher surprisals would be recorded if the head syntax nodes in Fig. 1 were R and I instead of I and P, or if the lower syntax nodes had illegal pairings within the same constituent in Fig. 1.

Sometimes the nodes in the tree could terminate with no further divisions. Such terminations would have the panels as leaves in the tree, as in Fig. 1. An entire narrative arc can thus be represented using a tree-like structure to create a Probabilistic Context Free Grammar in visual narratives, which can then be used with the existing natural language processing libraries such as NLTK (Bird, Klein, & Loper, 2009). Furthermore, this framework also allows incorporating representations of panel-level semantic information as described by the situational changes (see below). Thus, Earley parsers can incorporate sophisticated hierarchical grammar rules alongside semantic information to investigate hierarchical processing in visual narratives. We implemented two versions of Earley parsers, one with semantic information, and one without. In the version without semantic information, narrative categories were treated as the terminal nodes in the tree. This was done to systematically evaluate the effects of narrative grammar and semantics within the framework of the parser. Earley parser was programmed in Python using custom libraries available on the osf repository <https://osf.io/s2h5x/>.

8.1.1. *Incorporating panel-level visual semantic information*

Semantic information was represented using situational changes between panels. For each panel in the comic strip, we used the same coding scheme from Experiment 1, where changes in character, location, and causality information were used as a proxy for our features of

the semantic content. Similar encoding schemes have been used elsewhere in the discourse processing literature (Frank, Koppen, Noordman, & Vonk, 2003, 2009; Venhuizen, Crocker, & Brouwer, 2019). These changes are coded as $[0, 0.5, 1]$ —where 0 is no change between the preceding and the current panel, 0.5 is a partial change, and 1 is a full change associated with that feature. Thus, each panel in the comic strip can be summarized in terms of a unique vector of changes $[a, b, c]$ where a, b, c assume the values $\{0, 0.5, 1\}$ compared to the previous panel. The first panel in each strip always had the semantic changes coded as $\{1, 1, 1\}$ since it had the highest semantic change information. These vectors are a proxy for visual information in the panels, and therefore, can then be incorporated as the terminal nodes alongside the narrative grammar in the models of visual narrative comprehension. This information is then used to compute the probability of observing each panel's contents given the underlying tree structure and the prior panel contents. The computed probability is then used to compute the surprisals described in Eq. 1. A Python implementation of the Earley parser can be accessed via the Python notebook (see Experiment 2.ipynb) at <https://osf.io/s2h5x/>

8.2. Hidden Markov models

HMMs are a different class of models that are widely used in sequence predictions. Unlike the Earley parser, these are not top-down models. They make predictions about upcoming units in a sequence by learning the underlying distribution of the sequences in the data. Specifically, they assume a finite number of hidden states that result in the production of the observed sequences. An HMM sequence prediction is comprised of three components: (1) the starting probability of the states, (2) a matrix of state transition probabilities, and (3) emission probabilities. Start probabilities reflect prior knowledge—that is, which state is most likely to start the sequence. Transition probabilities guide the evolution of the sequence—that is, they reflect the probability of moving into another state given the current one. The transition probability matrix is assumed to be Markovian, that is, it only considers the immediate previous state and the current state transitions, ignoring the previous state transition history. Finally, the emission probabilities reflect the probability of observing a unit in the sequence given the current state. Like the Earley parser, the HMM framework could also incorporate panel-level semantic information alongside the narrative categories.

Panel-level information as described by the semantic changes was treated as an observed unit for each panel within each sequence, while the narrative categories were modeled as the underlying latent states. The start probabilities, therefore, describe the starting probability of the states—that is, the probability with which a given narrative category (e.g., E, I, P, R) occurs at the start of a sequence. Transition probabilities describe the probability of the next narrative state in a sequence given the current state. Finally, the emission probabilities described the probability of observing the panel-level information given the current state. See Figs. S1 and S2 for emission probabilities. The log-likelihood of observing a panel's specific content information is obtained by multiplying the state initial probability matrix with the transition and emission probability matrices using the forward and backward operations. The surprisals for HMMs were computed by plugging in the log-likelihoods into the surprisal equation—see Eq. 1. HMMs do not assume any recursion-based grammar rules unlike Earley parsers, and

therefore, do not make predictions based on tree-like structures. However, they do provide a framework to incorporate and model the visual narrative categories and semantic information in parallel. We implemented two versions of HMMs, one with semantic change information as the observed units, and one without. The latter version had narrative categories as both the hidden and observed units, while the former version had semantic changes as the observed units. This manipulation allowed us to systematically manipulate narrative categories and semantic change effects within the HMM framework. HMMs were implemented using the “hmmlearn – version 0.2.6” Python library. The code can be accessed via the Python notebook at <https://osf.io/s2h5x/>.

8.3. *N-gram models*

N-gram models compute the probability of observing a current unit in the sequence using previously observed n units in the sequence. Unlike HMMs and Earley parser, n-grams do not contain any latent states nor a hierarchical grammar, and therefore, cannot model both the semantic and syntax information in parallel. They rely only on the observed data in a sequence to compute the underlying statistical regularities either at the narrative category or the semantic levels. We considered the n-grams as a control condition for comparison alongside the Earley parser and HMMs. The n-gram surprisal for a panel w_i in the sequence is defined as $s(w_i) = p(w_i | w_{i-1}, \dots, w_{i-n-1}) = p(w_{i-1}, \dots, w_{i-n-1}) / p(w_{i-2}, \dots, w_{i-n-1})$ following the Bayes chain rule. We used bigrams ($n = 2$) and trigrams ($n = 3$) to model visual narrative comprehension. In these models, the input w_i comprised either semantic change or narrative category information but not both. Thus, separate n-gram models ($n = 2$ and 3) were fit for both semantic change and narrative category information, respectively. Henceforth, these will be referred to as the semantics n-grams and narrative category n-grams, respectively.

8.4. *Stimuli*

One hundred and twenty comic strips from Experiment 1 were used in this study. The panel-level information was coded for each panel as described above (see the section about incorporating panel-level visual information). Overall, the content in the 720 panels across 120 comic strips was described using 23 unique vectors representing collocations of changes in location, number of characters, and causality between panels (see Fig. 6 for a histogram plot).

The syntax trees were then generated for each of the 120 comic strips using VNG (Cohn, 2015). Both authors collaboratively generated the syntax trees following the rules and diagnostic tests of VNG. Furthermore, these trees were also validated by comparing them to an earlier empirical study where participants were asked to make a single segmentation into constituent sequences (Cohn et al., 2014). Overall, there were about 60 unique syntax trees for the 120 strips (see Fig. 7 for a histogram plot). The syntax trees for the 120 strips are available on the osf repository for further reference (<https://osf.io/s2h5x/>). Each tree comprised the rules of narrative grammar in addition to incorporating the panel-level change information in the terminal nodes. The entire 120 trees were then used to create the tree bank. Pairwise pairings of the narrative categories (e.g., E-I, I-P, etc.) were obtained from the same tree bank

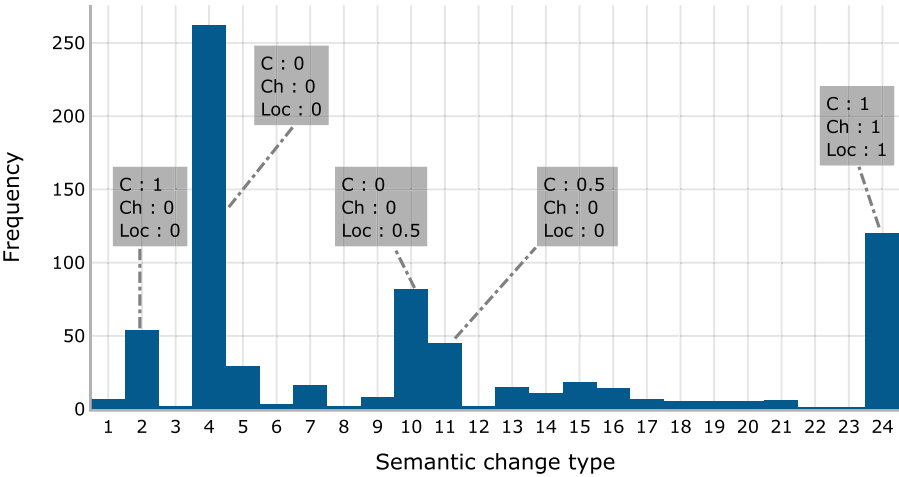


Fig. 6. Semantic change distribution across the 120 comic strips. Each change type is encoded as a three-dimensional vector tracking the Causal change (C), Character change (Ch), and Location change (Loc). The values along each of the dimensions correspond to {0, 0.5, 1} denoting a no-change, partial change, and a full change, respectively. The y-axis reflects the frequency of occurrence of each semantic type vector in the corpus of the 120 comic strips.

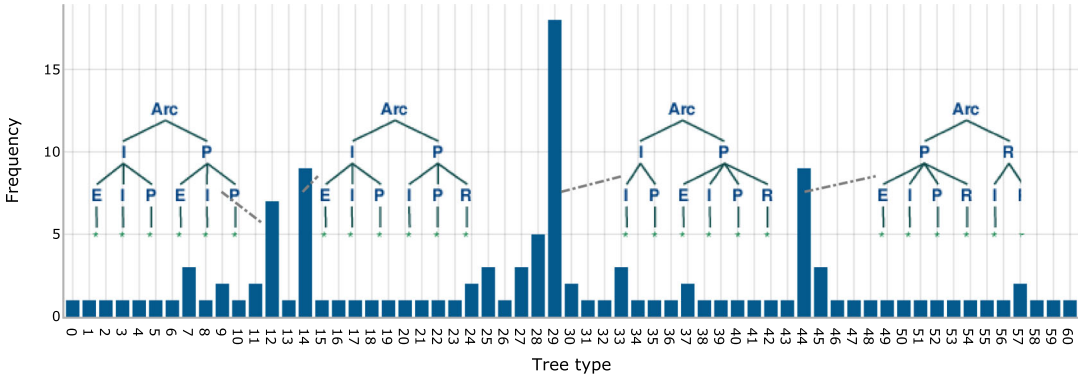


Fig. 7. Narrative grammar tree distribution across the 120 comic strips. Overall, there were 60 unique narrative trees. The asterisk (*) at the end of each tree signifies situational change information as shown in Fig. 6.

to compute the start and state transition probability matrices for the HMMs and the syntax n-grams models. Overall, there were 19 unique narrative category bigram pairs (see Fig. 3 for a histogram plot). Finally, the panel-level semantic changes were paired with their respective narrative categories to compute the emission probabilities for the HMM.

8.5. Procedure

We tested the performance of the Earley parser, HMM, and n-grams on the segmentation paradigm presented in Experiment 1. Each of the 120 comic strip sequences was input to

these models. To generalize both the models' performance on unseen data, the models were trained on 90%¹ of the comic strips randomly sampled from the total 120 strips excluding the currently modeled strip. This resulted in the creation of a separate tree bank, HMM probability matrices, and n-grams for each of the 120 strips. This procedure was further repeated 100 times for each strip, wherein we computed the log-likelihood of observing the next panel given the sequence, and subsequently the associated surprisal for each panel. Surprisals were then averaged across the 100 runs for each panel in the 120 strips to generate an estimated average surprisal for each panel in the 120 strips. The averaged surprisals were then ranked in descending order to obtain boundary preferences. Higher ranks would imply a likely boundary. Furthermore, to characterize the relationship between surprisals and the constituent boundaries, the distance of the panel from the boundary was also incorporated into our analysis. In each comic sequence, a panel from the previous constituent before the boundary was coded as -1 , and the panel following the boundary in the subsequent constituent was coded as $+1$, and so on. For example, in a 2-4 comic strip—where the constituent boundary occurs after the second panel, the panel positions relative to the boundary would be coded as -2 , -1 , 1 , 2 , 3 , 4 , respectively, implying that the panels were at a distance of -2 to $+4$ from the boundary. Depending on the constituent pattern type (e.g., 2-4, 3-3, 4-2), the panel distance from the constituent boundary varied from -4 to $+4$. Accordingly, we should expect a higher surprisal from panel -1 to $+1$, that is, the panel before and after the boundary.

9. Results

There were 120 strips with six panels each. The Earley parser, HMM, semantics n-gram, and the syntax n-gram models were used to predict a surprisal score for each panel. Thus, a total of 720 surprisals resulted each from the Earley parser, HMM, and both the n-gram implementations. The following surprisal scores were observed across the different models: Earley surprisals with narrative grammar and semantics ($M = 3.40$ bits, $SD = 1.9$), Earley surprisals with only narrative grammar ($M = 1.74$ bits, $SD = 1.51$), HMM surprisals with narrative grammar and semantics ($M = 3.05$, $SD = 2.77$), HMM surprisals with only narrative grammar information ($M = 1.35$ bits, $SD = 2.82$), and the n-grams (semantic trigram: $M = 3.65$, $SD = 1.26$; semantic bigrams: $M = 3.46$, $SD = 1.45$; narrative category trigrams: $M = 1.56$, $SD = 1.24$; narrative category bigrams: $M = 1.56$, $SD = 1.40$)—see Fig. 8 top row. Additionally, surprisals for the panels in each comic strip were normalized by the maximum recorded surprisal in the strip for comparison.

Each computational psycholinguistics model differed either in the inputs or the underlying mechanism thus making it uninterpretable to statistically compare their performance by including them as predictor variables in a single statistical model. We, therefore, fit separate linear mixed effects models to compare the generated surprisals against boundary agreement scores. In each statistical model, the participants' boundary agreement score was the dependent variable, and the generated surprisal scores were the predictor variables. Like the Experiment 1 analysis, the boundary number was modeled as a random effect at the intercept level in all the models. Further, given the five boundary agreement scores, and the six

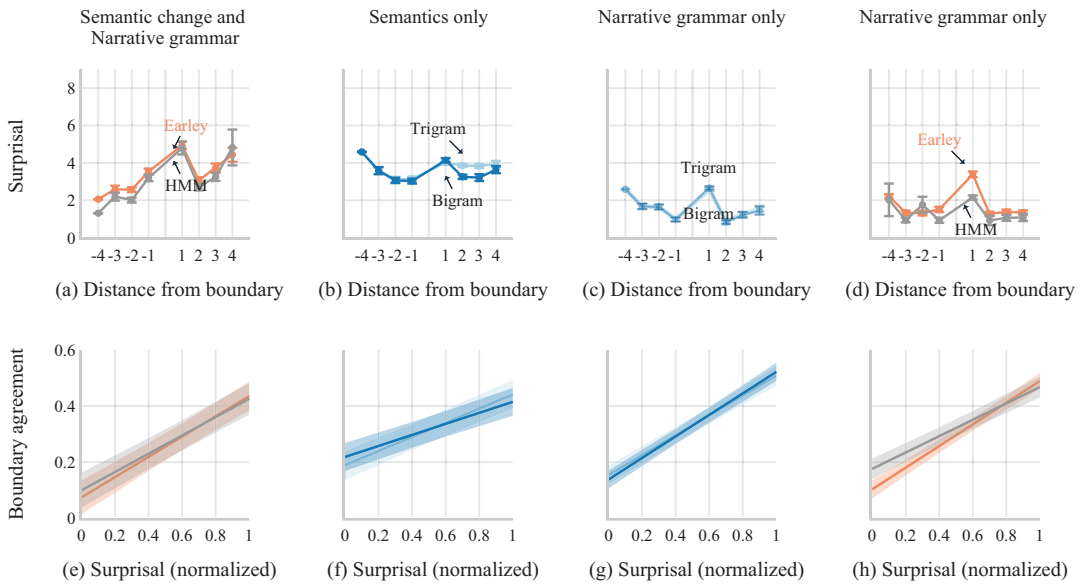


Fig. 8. Model performance comparison. The titles indicate the kind of information the computational psycholinguistics models had access to. “Semantics” refers to the situational changes such as changes in space, location, and causality, and “Narrative grammar” refers to the E, I, P, R, and the associated pairings/grammar rules depending on the model. The top panels plot the model surprisals as a function of distance from the boundary. The bottom panels plot their respective normalized surprisals against the boundary agreement. Error bars and shaded lines represent standard error and 95% CI, respectively.

computed surprisals in each strip, we removed the first generated surprisal from the models to match it with the boundary agreement scores. These analyses revealed that boundary agreement was significantly predicted by all the models. The model estimates are summarized in Table 2. Given that each of the models differed internally in the number of parameters, we only report the Akaike Information Criterion (AIC) scores for each model. When comparing various statistical models, AIC score is a useful metric to draw inferences on the goodness of the statistical model. Defined as the negative log-likelihood of the statistical model, a lower AIC score is an indicator of the goodness of the model fit when comparing various statistical models. On average, both the Earley parsers—with narrative grammar and semantics, and the narrative grammar alone—had lower AIC scores compared to the HMMs. The n-gram models also had lower AIC scores. However, see the General Discussion for more information regarding interpreting the n-gram models as a framework for visual narrative comprehension. Overall, the surprisal scores generated by all the models significantly predicted the boundary agreement score. The model prediction plots are displayed in the bottom row of Fig. 8.

We next evaluated whether the generated surprisals were significantly different for the panels immediately before and after the theoretically identified boundaries. Paired *t*-tests also revealed significantly higher surprisals for the panel after a boundary compared to the panel before for all the models—Earley narrative grammar + semantics: ($t(119) = 6.18, p < .001$); HMM narrative grammar + semantics: ($t(119) = 6.35, p < .001$); Semantics-only trigram

Table 2
Comparing GLME model estimates

Model	Predictors	Boundary agreement			AIC	Random effects, <i>SD</i>
		Estimate	<i>CI</i>	<i>p</i>		<i>Boundary number</i>
1	Intercept	0.07	[−0.04, 0.19]	.262	−39.16	0.09
	Earley parser with narrative grammar and semantics	0.36	[0.25, 0.46]	<.001		—
2	Intercept	0.34	[0.17, 0.50]	.004	−38.71	0.10
	HMM with narrative grammar and semantics	0.32	[0.22, 0.42]	<.001		—
3	Intercept	0.18	[0.08, 0.28]	.021	−34.09	0.08
	Semantics bigram	0.25	[0.17, 0.33]	<.001		—
4	Intercept	0.15	[0.08, 0.22]	.006	−82.58	0.05
	Narrative grammar bigram	0.35	[0.28, 0.42]	<.001		—
5	Intercept	0.21	[0.11, 0.31]	.01	−19.80	−0.08
	Semantics trigram	0.19	[0.11, 0.27]	<.001		—
6	Intercept	0.13	[0.07, 0.20]	.006	−91.23	0.04
	Narrative grammar trigram	0.38	[0.31, 0.45]	<.001		—
7	Intercept	0.10	[0.03, 0.16]	.01	−71.38	0.03
	Earley parser with only narrative grammar	0.38	[0.30, 0.46]	<.001		—
8	Intercept	0.17	[0.10, 0.24]	.005	−48.26	0.05
	HMM with only narrative grammar	0.29	[0.21, 0.36]	<.001		—

Note. Significant effects are highlighted in bold.

($t(119) = 6.02, p < .001$); Semantics-only bigram ($t(119) = 6.70, p < .001$); Narrative grammar-only trigram ($t(119) = 14.68, p < .001$); Narrative grammar-only bigram ($t(119) = 13.72, p < .001$); Earley Narrative grammar-only ($t(119) = 11.3, p < .001$); HMM Narrative grammar-only ($t(119) = 9.61, p < .001$). See Fig. 9 for more information.

Finally, we computed the frequency with which all the models rated a boundary between two constituents as their first preferred boundary, as opposed to boundaries elsewhere in the comic strip. This was done by ranking the computed surprisals generated by all the models. Overall, all the models predominantly ranked the boundary between two constituents (BC) as the first boundary compared to within constituents (WC), except for the semantics trigram model—Earley (BC = 42.5%, WC = 11.66%); HMM (BC = 45.83%, WC = 10%); Semantics trigram (BC = 16.667%, WC = 10%); Semantics bigram (BC = 33.3%, WC = 16.66%); Syntax trigram (BC = 55.8%, WC = 7.5%); Syntax bigram (BC = 55%, WC = 6.66%); Earley narrative grammar only (BC = 56.6%, WC = 13.3%); HMM narrative grammar only (BC = 43.33%, WC = 13.33%). See Fig. 10 for more information.

10. Discussion

This experiment used a computational framework to investigate the role of hierarchical grammar in visual narrative comprehension. Specifically, we used a probabilistic

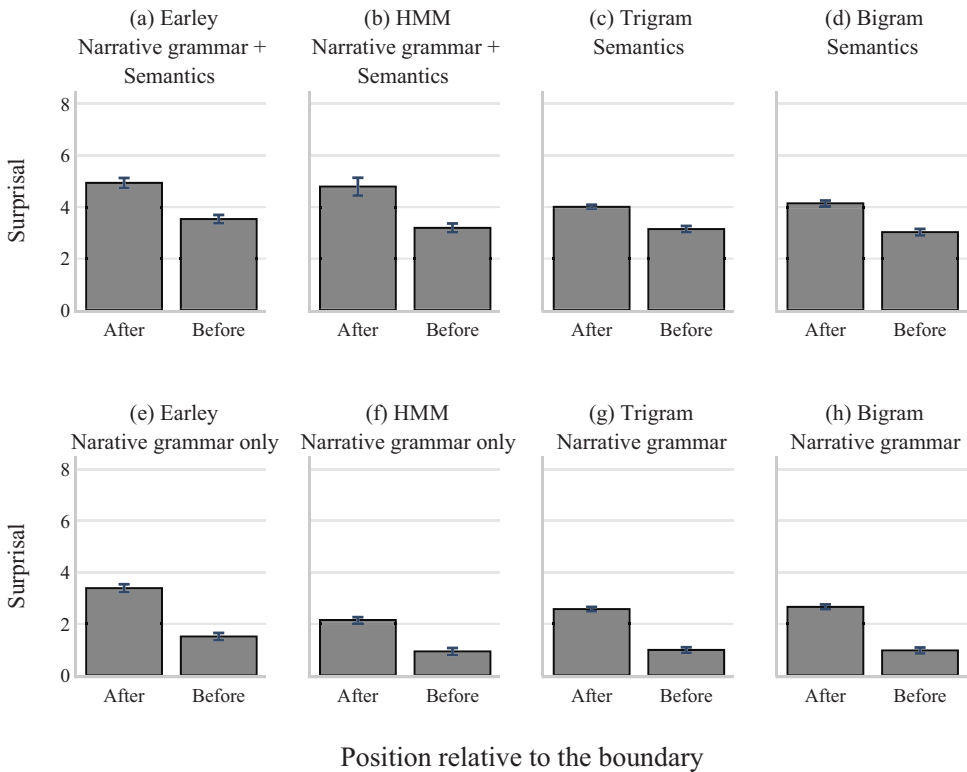


Fig. 9. Model surprisals recorded at constituent boundaries.

Earley parser, an HMM, and n-grams to model segmentation preferences in visual narrative sequences. All the models recorded surprisals—that is, how likely is the next unit expected given the history—for the upcoming panel in the visual narrative sequence conditioned on the previous input. Earley parser used a hierarchical VNG along with semantic change information. The HMMs used linear pairings of the narrative categories as the latent states and semantic change information as the observed states. N-grams on the other hand used either semantic change information or the narrative category information, but not both. The locations where the surprisals were highest in the sequence were considered as the models’ preferred boundary.

Overall, all the models predicted boundary agreement score as seen in Figs. 8e–h and Table 2. Furthermore, all the models also recorded high surprisals at the theoretically hypothesized constituent boundary as determined by VNG (see Figs. 8a–d and 9). Finally, we also analyzed the boundary preferences by ranking the generated surprisals for all the models for the segmentation task. Here, we found that all the models except the semantics trigram, consistently ranked the boundary between two constituents (BC) higher compared to the boundaries within a constituent (WC)—see Fig. 10. Note that the chance level here is at 20%—that is, one out of five possible locations. All the models except the semantics trigram ranked

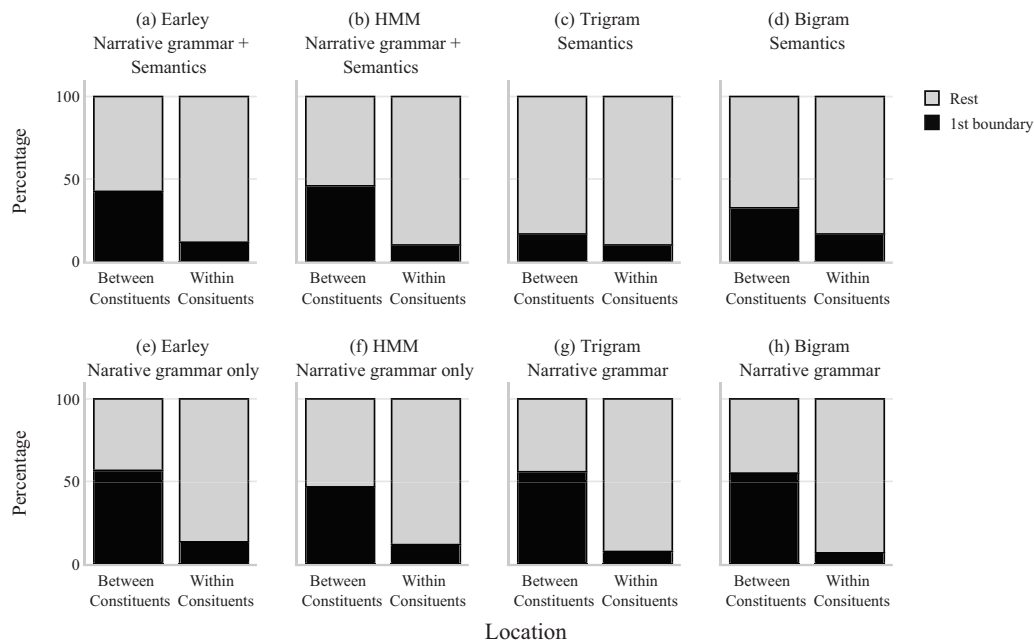


Fig. 10. Boundary preferences. Dark regions indicate the percentage of time the constituent boundary was their first preferred boundary. Light regions indicate the opposite.

the boundaries between constituents (BC) as their first preference—significantly above 20%. These results establish that participant segmentation preferences in visual narratives can be captured reliably using a computational framework.

Interestingly, both Earley parser and HMM recorded higher surprisals for panels 4 units after the boundary—see Fig. 8a. This is consistent with longer fixations and panel viewing times observed to the last panel in visual narrative sequences (Cohn, 2012; Foulsham, Wybrow, & Cohn, 2016). It is possible that panel-level changes such as changes in characters, locations, and so on might be driving the high surprisals in the last panels. The segmentation analysis using the syntax-only versions of Earley parser and HMM showed a significant drop in surprisals for the same panels that are 4 units away from the boundary. This analysis resulting in the drop in the surprisals did not use any semantic change information, thus confirming the role of semantic changes in driving high surprisals in the final panel in a sequence (see Fig. 8d). Future work could benefit from a more detailed investigation of the interaction between semantics and narrative structures during comics comprehension. Specifically, it would be interesting to characterize the predictions of both semantics and narrative structures for longer sequences.

A key takeaway from this analysis is that the surprisals generated using a computational framework were significant predictors of the boundary agreement score. It should be noted that the input to these models included the VNG and the semantic change information. These factors were individually shown to predict boundary agreement in Experiment 1 (see Fig. 5).

Similarly, n-grams individually incorporating semantic and syntactic information were also significant predictors of the boundary agreement. These results are in line with the findings of Cohn and Bender (2017).

11. General discussion

This study investigated whether visual narratives use a hierarchical narrative grammar. Previous work on visual narrative comprehension has demonstrated that participants group comic panels into constituents during visual narrative comprehension (Cohn et al., 2014; Cohn, 2012; Gernsbacher, 1985; Hagmann & Cohn, 2016). However, it is unclear how these grouping mechanisms account for nonadjacent relationships between panels. VNG addresses this issue by proposing grammatical properties that combine with the underlying semantics to convey a meaningful story. Across two experiments, we provided empirical evidence for the presence of a hierarchical grouping involving grammar in visual narrative comprehension.

Experiment 1 reanalyzed segmentation data from Cohn and Bender (2017), which investigated constituent boundaries in visual sequences by actively asking participants to mark the perceived boundaries. We found that participants significantly ranked the boundary between two constituents as their first preferred boundary as opposed to the boundaries elsewhere in the comic strip. Furthermore, participants' boundary agreement scores were predicted by both narrative category bigrams, as well as semantic changes such as changes in characters and location, thus replicating the findings of Cohn and Bender (2017).

Nevertheless, within these semantic changes, causal changes were not significant predictors of the boundary agreement—see Table 1. This is an interesting result because causal changes are a significant factor in determining segmentation preferences (Mathis & Papafragou, 2022; Riedl & Young, 2006, 2010; Suh & Trabasso, 1993; Trabasso, Van Den Broek, & Suh, 1989; Zacks et al., 2009), though not all the time (Kurby & Zacks, 2012; M. E. Smith, Kurby, & Bailey, 2023). One possible reason is that while causation has been shown to be a driver of “narrative” understanding at a situational level, this does not mean it is captured by the narrative grammar. As demonstrated in Cohn and Bender (2017), segmentation here is more predicted by narrative grammar than situational dimensions, while empirical evidence has shown the separability of situational dimensions like causation and the narrative grammar (see Cohn, 2020), including cases where the narrative grammar maintains structural legality despite no situational coherence, including causal structure (Cohn et al., 2012). Thus, the lack of influence of causal structure in our findings reinforces the separability of these situational dimensions like causation and the parallel narrative grammar.

In addition, in the correspondence between situational and narrative structures, causal relations are predicted to shift *within* narrative constituents (i.e., where actions tied to a common event take place)—more than they would be predicted to mark shifts *between* constituents (i.e., prototypically event boundaries) (Cohn & Bender, 2017). Such results again advocate for segmentation tasks to be more sensitive to the actual narrative structure of the discourse rather than situational dimensions alone.

Experiment 2 then investigated the role of a hierarchical grammar using a computational framework. Specifically, different computational psycholinguistics models of sentence

comprehension—probabilistic Earley parsers (Hale, 2001; Stolcke, 1994), HMMs (Juang & Rabiner, 1991; Rabiner & Juang, 1986), and n-grams were used to investigate the behavioral segmentation data from Experiment 1. All the models computed surprisals—the probability of expecting a particular panel next in the sequence given the current observations. Surprisals have been shown to capture word processing difficulty, and thus the reading time in the psycholinguistics literature (Hale, 2001, 2016; Levy, 2008; N. J. Smith & Levy, 2013). It has been widely shown both for language and in the literature on visual narrative comprehension that reading times significantly increase at the beginning of a constituent (Cohn et al., 2014; Cohn et al., 2012; Levy, 2008; Linzen et al., 2016). Thus, if surprisals are a cognitive marker for the underlying comprehension process in visual narratives, we reasoned they should be able to inform us about the location of preferred narrative constituent boundaries. To obtain the boundary preferences, the computed surprisals of both the Earley parser, the HMMs, and the n-grams were ranked and compared with that of the participant boundary preferences from the segmentation task in Experiment 1.

The surprisals generated from all the different models were significant predictors of the boundary agreement. These results reinforce the findings that semantics and narrative grammar are crucial for narrative comprehension. This can also be seen from the results of Experiment 1, and the findings of Cohn and Bender (2017). However, not all the models implemented here adhere closely to the theoretical frameworks. For example, note that the n-grams do not explicitly model narrative grammar and semantics in parallel, unlike the Earley parser and the HMM. Theories of visual narrative comprehension have proposed that both semantic and syntactic structures are crucial for comprehension (Cohn, 2020; Loschky et al., 2018, 2020; Zwaan & Radvansky, 1998), with neural evidence backing this claim (Cohn et al., 2012). Furthermore, modeling the semantic and syntactic structures in parallel involves accounting for the statistical transitions both within the syntactic structures (e.g., the probability with which the narrative category “E” transitions into an “I”) and also between the syntactic and semantic structures (e.g., the probability with which a narrative category results in a particular semantic change). It should be noted that n-grams cannot explicitly incorporate such parallelism. Therefore, any attempt to model both the narrative syntax and semantics information in parallel should include a nonflattened dual-level comprehension mechanism at the very least. We, therefore, restrict the rest of our discussion to the Earley parser and the HMM with narrative syntax and semantics information.

Both the Earley parser and the HMM with narrative syntax and semantics information predicted the segmentation agreement score. Both the models incorporated a hierarchical grammar that can effectively combine the narrative category and semantic change information. However, the Earley parser combined this information using principles of recursion, unlike the HMMs which modeled statistical regularities between the narrative grammar and semantics information. While the two frameworks are useful in capturing intricate statistical regularities in the joint probability space of narrative grammar and semantics, the mechanistical differences between the two could provide crucial insights into visual narrative comprehension. For example, it would be interesting to characterize the performance of the two models for longer sequences—a current limitation of this study where strip sequences of six panels each were tested. Furthermore, all the strips used in the study had a linear narrative structure—where


there were no subnarratives occurring inside the main narrative. Recent work has shown that participants have the ability to track nonlinearly structured narratives presented in the form of intentional situational incongruities, and that the similarity between the two narratives affects the narrative comprehension (Chan, Foy, & Magliano, 2018; Klomberg, Schilperord, & Cohn, 2024). Therefore, comparing the performance of the Earley parser and HMM on nonlinear narrative sequences could provide crucial insights into understanding the mechanistic underpinnings of visual narrative comprehension. Future work could benefit from a more detailed investigation on this front.

A key implication of this result has to do with the importance of hierarchical organization in our visual narrative comprehension. Though many discourse approaches acknowledge boundaries and hierarchic structure, these theories have focused on the changes that occur between adjacent panels such as between characters, spatial locations, causation, and connections to broader semantic associations (Magliano et al., 2001; Magliano & Zacks, 2011; McCloud, 1993; Saraceni, 2001; Zwaan et al., 1995; Zwaan & Radvansky, 1998). These linear relationships have also been used to describe comprehension in film viewing (Magliano et al., 2001; Magliano & Zacks, 2011; Zacks et al., 2009), and other approaches to visual narratives have also focused on pairwise relationships between the panels (McCloud, 1993). While it is true that these semantic relationships modulate how event boundaries are perceived, in the presence of broader thematic roles as directed by the narrative grammar, our narrative comprehension could benefit from combining both the narrative grammar and the semantic relationships—as demonstrated by both the Earley parser and the HMM models. A more detailed characterization of the performance of these computational models against human behavior could provide crucial insights into understanding the mechanistic nature of visual narrative comprehension.

Acknowledgments

We are grateful for the initial discussions with Tal Linzen, Suhas Arehalli, and Yash Kumar Lal while the first author was a graduate student at Johns Hopkins University. We also thank Jeffrey M. Zacks, Lester Loschky, and two other anonymous reviewers for their helpful comments about improving the manuscript.

Open Research Badges

 This article has earned Open Data and Open Materials badges. Data and Materials are available at <https://osf.io/s2h5x/>.

Note

- 1 We also repeated the analysis with an 80–20% split between training and the test data. The results did not change. See Fig. S3 for more details.

References

- Baldassano, C., Chen, J., Zadbood, A., Pillow, J. W., Hasson, U., & Norman, K. A. (2017). Discovering event structure in continuous narrative perception and memory. *Neuron*, 95(3), 709–721.e5.
- Baldassano, C., Hasson, U., & Norman, K. A. (2018). Representation of real-world event schemas during narrative perception. *Journal of Neuroscience*, 38(45), 9689–9699.
- Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: Analyzing text with the natural language toolkit*. O'Reilly Media, Inc.
- Bransford, J. D., & Johnson, M. K. (1972). Contextual prerequisites for understanding: Some investigations of comprehension and recall. *Journal of Verbal Learning and Verbal Behavior*, 11(6), 717–726.
- Brown, G., Brown, G. D., Brown, G. R., Yule, G., & Gillian, B. (1983). *Discourse analysis*. Cambridge University Press.
- Chan, G. C., Foy, J. E., & Magliano, J. P. (2018). Factors that affect crossover between multiple worlds within a narrative. *Discourse Processes*, 55(8), 666–685.
- Chomsky, N. (1956). Three models for the description of language. *IRE Transactions on Information Theory*, 2(3), 113–124.
- Chomsky, N. (1957). Logical structure in language. *Journal of the American Society for Information Science*, 8(4), 284.
- Chomsky, N. (2009). *Syntactic structures*. De Gruyter Mouton.
- Chomsky, N. (2013). Problems of projection. *Lingua*, 130, 33–49.
- Cohn, N. (2012). *Structure, meaning, and constituency in visual narrative comprehension* [PhD Thesis]. Tufts University.
- Cohn, N. (2013a). *The visual language of comics: Introduction to the structure and cognition of sequential images*. A&C Black.
- Cohn, N. (2013b). Visual narrative structure. *Cognitive Science*, 37(3), 413–452.
- Cohn, N. (2014). The architecture of visual narrative comprehension: The interaction of narrative structure and page layout in understanding comics. *Frontiers in Psychology*, 5, 680. <https://doi.org/10.3389/fpsyg.2014.00680>
- Cohn, N. (2015). How to analyze visual narratives: A tutorial in Visual Narrative Grammar. Retrieved from http://Www.Visuallanguagelab.Com/P/VNG_Tutorial.Pdf.
- Cohn, N. (2020). Your brain on comics: A cognitive model of visual narrative comprehension. *Topics in Cognitive Science*, 12(1), 352–386.
- Cohn, N. (2023). *The patterns of comics: Visual languages of comics from Asia, Europe, and North America*. Bloomsbury Publishing.
- Cohn, N., & Bender, P. (2017). Drawing the line between constituent structure and coherence relations in visual narratives. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 43(2), 289.
- 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.
- Cohn, N., & Kutas, M. (2017). What's your neural function, visual narrative conjunction? Grammar, meaning, and fluency in sequential image processing. *Cognitive Research: Principles and Implications*, 2(1), 27.
- Cohn, N., & Paczynski, M. (2013). Prediction, events, and the advantage of agents: The processing of semantic roles in visual narrative. *Cognitive Psychology*, 67(3), 73–97.
- 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), 1–38.
- Cuddon, J. A., & Habib, R. (2013). *A Dictionary of literary terms and literary theory* (5th ed). Wiley Blackwell.
- Dehaene, S., Meyniel, F., Wacongne, C., Wang, L., & Pallier, C. (2015). The neural representation of sequences: From transition probabilities to algebraic patterns and linguistic trees. *Neuron*, 88(1), 2–19.
- Fodor, J. A., & Bever, T. G. (1965). The psychological reality of linguistic segments. *Journal of Verbal Learning and Verbal Behavior*, 4(5), 414–420.

- Fodor, J., Bever, A., & Garrett, T. (1974). *The psychology of language: An introduction to psycholinguistics and generative grammar*.
- Foulsham, T., Wybrow, D., & Cohn, N. (2016). Reading without words: Eye movements in the comprehension of comic strips. *Applied Cognitive Psychology*, 30(4), 566–579.
- Frank, S. L., Haselager, W. F. G., & Van Rooij, I. (2009). Connectionist semantic systematicity. *Cognition*, 110(3), 358–379.
- Frank, S. L., Koppen, M., Noordman, L. G. M., & Vonk, W. (2003). Modeling knowledge-based inferences in story comprehension. *Cognitive Science*, 27(6), 875–910.
- Franklin, N. T., Norman, K. A., Ranganath, C., Zacks, J. M., & Gershman, S. J. (2020). Structured event memory: A neuro-symbolic model of event cognition. *Psychological Review*, 127(3), 327–361.
- Friederici, A. D. (2011). The brain basis of language processing: From structure to function. *Physiological Reviews*, 91(4), 1357–1392.
- Gernsbacher, M. A. (1985). Surface information loss in comprehension. *Cognitive Psychology*, 17(3), 324–363.
- Gershman, S. J., Radulescu, A., Norman, K. A., & Niv, Y. (2014). Statistical computations underlying the dynamics of memory updating. *PLoS Computational Biology*, 10(11), e1003939.
- Hagmann, C. E., & Cohn, N. (2016). The pieces fit: Constituent structure and global coherence of visual narrative in RSVP. *Acta Psychologica*, 164, 157–164.
- Hagoort, P., & Indefrey, P. (2014). The neurobiology of language beyond single words. *Annual Review of Neuroscience*, 37(1), 347–362.
- Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.
- Hale, J. (2016). Information-theoretical complexity metrics. *Language and Linguistics Compass*, 10(9), 397–412.
- Halliday, M. A., Hasan, R., & Hasan, R. (1985). *Language, text and context*. Geelong: Deakin University Press.
- Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve? *Science*, 298(5598), 1569–1579.
- Holquist, M. (2003). *Dialogism*. Routledge.
- Huff, M., Meitz, T. G., & Papenmeier, F. (2014). Changes in situation models modulate processes of event perception in audiovisual narratives. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40(5), 1377.
- Jackendoff, R. (2009). Parallels and nonparallels between language and music. *Music Perception*, 26(3), 195–204.
- Jackendoff, R., & Lerdahl, F. (2006). The capacity for music: What is it, and what's special about it? *Cognition*, 100(1), 33–72.
- Juang, B. H., & Rabiner, L. R. (1991). Hidden Markov models for speech recognition. *Technometrics*, 33(3), 251–272.
- Jurafsky, D., Wooters, C., Segal, J., Stolcke, A., Fosler, E., Tajchaman, G., & Morgan, N. (1995). Using a stochastic context-free grammar as a language model for speech recognition. *1995 International Conference on Acoustics, Speech, and Signal Processing*, 1, 189–192.
- Klomberg, B., Schilperoord, J., & Cohn, N. (2024). Constructing domains in visual narratives: Structural patterns of incongruity resolution. *Journal of Comparative Literature and Aesthetics*, 47(3), 37–55.
- Kurby, C. A., & Zacks, J. M. (2012). Starting from scratch and building brick by brick in comprehension. *Memory & Cognition*, 40(5), 812–826.
- Laubrock, J., & Dunst, A. (2020). Computational approaches to comics analysis. *Topics in Cognitive Science*, 12(1), 274–310.
- Lerdahl, F., & Jackendoff, R. (1983). An overview of hierarchical structure in music. *Music Perception*, 1(2), 229–252. <https://doi.org/10.2307/40285257>.
- Lerdahl, F., & Jackendoff, R. S. (1996). *A generative theory of tonal music*. MIT Press.
- Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106(3), 1126–1177.
- Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4, 521–535.

- Loschky, L. C., Hutson, J. P., Smith, M. E., Smith, T. J., & Magliano, J. P. (2018). Viewing static visual narratives through the lens of the scene perception and event comprehension theory (SPECT). In A. Dunst, J. Laubrock, & J. Wildfeuer (Eds.), *Empirical comics research: Digital, Multimodal, and Cognitive Methods* (pp. 217–238). New York, NY: Routledge.
- Loschky, L. C., Larson, A. M., Smith, T. J., & Magliano, J. P. (2020). The scene perception & event comprehension theory (SPECT) applied to visual narratives. *Topics in Cognitive Science*, 12(1), 311–351.
- Magliano, J. P., Miller, J., & Zwaan, R. A. (2001). Indexing space and time in film understanding. *Applied Cognitive Psychology*, 15(5), 533–545.
- Magliano, J. P., & Zacks, J. M. (2011). The impact of continuity editing in narrative film on event segmentation. *Cognitive Science*, 35(8), 1489–1517.
- Mandler, J. M., & Johnson, N. S. (1977). Remembrance of things parsed: Story structure and recall. *Cognitive Psychology*, 9(1), 111–151.
- Marcus, G. F. (2013). Evolution, memory, and the nature of syntactic representation. In *Birdsong, speech, and language: Exploring the evolution of mind and brain*.
- Martens, C., Cardona-Rivera, R. E., & Cohn, N. (2020). The visual narrative engine: A computational model of the visual narrative parallel architecture. In *Proceedings of the 8th Annual Conference on Advances in Cognitive Systems*.
- Mathis, A., & Papafragou, A. (2022). Agents' goals affect construal of event endpoints. *Journal of Memory and Language*, 127, 104373.
- McCloud, S. (1993). *Understanding comics: The invisible art*. Northampton, MA.
- Mura, K., Petersen, N., Huff, M., & Ghose, T. (2013). IBES: A tool for creating instructions based on event segmentation. *Frontiers in Psychology*, 4, 994.
- Newson, D. (1973). Attribution and the unit of perception of ongoing behavior. *Journal of Personality and Social Psychology*, 28(1), 28.
- Newson, D., & Engquist, G. (1976). The perceptual organization of ongoing behavior. *Journal of Experimental Social Psychology*, 12(5), 436–450.
- Pantaleo, S. (2004). The long, long way: Young children explore the Fabula and Syuzhet of shortcut. *Children's Literature in Education*, 35(1), 1–20.
- Pearce, M., & Rohrmeier, M. (2018). Musical syntax II: Empirical perspectives. In R. Bader (Ed.), *Springer handbook of systematic musicology* (pp. 487–505). Springer Handbooks. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-662-55004-5_26.
- Qi, S., Jia, B., Huang, S., Wei, P., & Zhu, S.-C. (2020). A generalized Earley parser for human activity parsing and prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(8), 2538–2554.
- Rabiner, L., & Juang, B. (1986). An introduction to hidden Markov models. *IEEE ASSP Magazine*, 3(1), 4–16.
- Reynolds, J. R., Zacks, J. M., & Braver, T. S. (2007). A computational model of event segmentation from perceptual prediction. *Cognitive Science*, 31(4), 613–643.
- Riedl, M. O., & Young, R. M. (2006). From linear story generation to branching story graphs. *IEEE Computer Graphics and Applications*, 26(3), 23–31.
- Riedl, M. O., & Young, R. M. (2010). Narrative planning: Balancing plot and character. *Journal of Artificial Intelligence Research*, 39, 217–268.
- Rohrmeier, M., & Pearce, M. (2018). Musical syntax I: Theoretical perspectives. In R. Bader (Ed.), *Springer handbook of systematic musicology* (pp. 473–486). Springer Handbooks. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-662-55004-5_25.
- Rumelhart, D. E. (1975). Notes on a schema for stories. In D. G. Bobrow, & A. Collins (Eds.), *Representation and understanding: Studies in cognitive science* (pp. 211–236). Academic Press. <https://doi.org/10.1016/B978-0-12-108550-6.50013-6>.
- Saraceni, M. (2001). Relatedness: Aspects of textual connectivity in comics. In J. Baetens (Ed.), *The graphic novel* (13 ed., pp. 167–180). (Symbolae Facultatis Litterarum Lovaniensis, series D: litteraria; No. 13). Leuven University Press.
- Schank, R. C., & Abelson, R. P. (1975). Scripts, plans, and knowledge. *IJCAI*, 75, 151–157.

- Schank, R. C., & Abelson, R. P. (2013). *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Psychology Press.
- Smith, M. E., Kurby, C. A., & Bailey, H. R. (2023). Events shape long-term memory for story information. *Discourse Processes*, 60(2), 141–161.
- Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3), 302–319.
- Sportiche, D., Koopman, H., & Stabler, E. (2013). *An introduction to syntactic analysis and theory*. John Wiley & Sons.
- Sprouse, J., & Schütze, C. (2020). Grammar and the use of data. In B. Aarts, J. Bowie, & G. Popova (Eds.), *The Oxford handbook of English grammar* (1st ed., pp. 40–58). Oxford University Press.
- Stolcke, A. (1994). An efficient probabilistic context-free parsing algorithm that computes prefix probabilities. *arXiv Preprint Cmp-Lg/9411029*.
- Suh, S. Y., & Trabasso, T. (1993). Inferences during reading: Converging evidence from discourse analysis, talk-aloud protocols, and recognition priming. *Journal of Memory and Language*, 32(3), 279–300.
- Trabasso, T., Van Den Broek, P., & Suh, S. Y. (1989). Logical necessity and transitivity of causal relations in stories. *Discourse Processes*, 12(1), 1–25.
- Tran, M.-V., Collier, N., Le, H. Q., Phi, V.-T., & Pham, T.-B. (2013). Exploring a probabilistic Earley parser for event composition in biomedical texts. In *Proceedings of the BioNLP Shared Task 2013 Workshop* (pp. 130–134).
- Van Dijk, T. A. (1977). Semantic macro-structures and knowledge frames in discourse comprehension. *Cognitive Processes in Comprehension*, 332, 3–31.
- Venhuizen, N. J., Crocker, M. W., & Brouwer, H. (2019). Expectation-based comprehension: Modeling the interaction of world knowledge and linguistic experience. *Discourse Processes*, 56(3), 229–255.
- Zacks, J. M., Speer, N. K., & Reynolds, J. R. (2009). Segmentation in reading and film comprehension. *Journal of Experimental Psychology: General*, 138(2), 307.
- Zacks, J. M., & Tversky, B. (2001). Event structure in perception and conception. *Psychological Bulletin*, 127(1), 3.
- Zwaan, R. A., Langston, M. C., & Graesser, A. C. (1995). The construction of situation models in narrative comprehension: An event-indexing model. *Psychological Science*, 6(5), 292–297.
- Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, 123(2), 162.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Supplemental Information