Journal of Pragmatics 127 (2018) 1–19

Contents lists available at ScienceDirect

Journal of Pragmatics

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

In defense of a "grammar" in the visual language of comics

Neil Cohn*

Tilburg Center for Cognition and Communication, Tilburg University, Tilburg, The Netherlands

ARTICLE INFO

Article history:
Received 27 January 2017
Received in revised form 1 January 2018
Accepted 5 January 2018

Keywords:
Visual language
Narrative grammar
Comics
Linguistics
Empirical research

ABSTRACT

Visual Language Theory (VLT) argues that the structure of drawn images is guided by similar cognitive principles as language, foremost a "narrative grammar" that guides the ways in which sequences of images convey meaning. Recent works have critiqued this linguistic orientation, such as Bateman and Wildfeuer's (2014) arguments that a grammar for sequential images is unnecessary. They assert that the notion of a grammar governing sequential images is problematic, and that the same information can be captured in a "discourse" based approach that dynamically updates meaningful information across juxtaposed images. This paper reviews these assertions, addresses their critiques about a grammar of sequential images, and then details the shortcomings of their own claims. Such discussion is directly grounded in the empirical evidence about how people comprehend sequences of images. In doing so, it reviews the assumptions and basic principles of the narrative grammar of the visual language used in comics, and it aims to demonstrate the empirical standards by which theories of comics' structure should adhere to.

1. Introduction

Visual Language Theory (VLT) outlines a framework for the structure and cognition of the "visual language" used in comics that is grounded in the methodologies of the linguistic and cognitive sciences. The keystone of this theory is the idea that sequential images use a "narrative grammar" that is structured and comprehended using similar mechanisms as grammars in verbal and signed languages. Given this perspective, it has not been without its doubters and detractors. Indeed, most prior approaches to visual narrative sequencing have characterized only the meaningful coherence changes between images (Gernsbacher, 1985; Magliano et al., 2001; McCloud, 1993; Saraceni, 2000, 2016; Stainbrook, 2003, 2016), and a grammatical model introduces complexity and machinery that may be perceived as unnecessary in comparison. For example, recent publications by Bateman and Wildfeuer (2014a, 2014b), henceforth "B&W", claim to outline the deficiencies of VLT and proffer a discourse-based model that can allegedly encompass what is offered by narrative grammar, only better.

In light of these concerns, this paper therefore has several aims. First, it will address the critiques made by B&W. This will show that their criticisms are unfounded, and that their alternate model not only cannot describe the empirical data about sequential image comprehension, but it fails to adequately describe phenomena in visual narratives like comics. Note that this discussion could extend to many approaches that have theorized about the meaningful relations between images (e.g.,

* Tilburg University, Tilburg center for Cognition and Communication (TiCC), P.O. Box 90153, 5000 LE Tilburg, The Netherlands.
E-mail address: neilcohn@visuallanguagelab.com.

https://doi.org/10.1016/j.pragma.2018.01.002
0378-2166/

McCloud, 1993; Saraceni, 2000, 2016; Stainbrook, 2003). However, the B&W model is perhaps the most theoretically sophisticated, and directly takes aim at VLT. Thus, it can serve as a good contrast.

Second, in doing so, this paper will serve as a review by which readers can familiarize themselves with the assumptions and methods of VLT. Specifically, this paper will focus on the structure and comprehension of *sequential* images.[1] Finally, this exercise can hopefully demonstrate and elucidate what type of standards are required for making claims about the structures used in the visual language in comics. Thus, comparison between theories is viewed as a fruitful exercise in the development of such a research field.

2. Visual Narrative Grammar

At the outset, we will review the basic principles of Visual Language Theory (VLT) and its sub-theory about sequential images, ***Visual Narrative Grammar*** (VNG). VLT argues that drawings use similar structural and cognitive principles as language (Cohn, 2013b). This involves encoding into memory systematic mappings between form (sounds, graphics) and meaning to create "lexical items" (words, images), which are then sequentially ordered using a grammatical system (syntax, narrative). This overall idea is guided by the assumption of "equivalence" that the mind/brain uses similar structures and principles across domains, except for those motivated from the differences in modalities themselves. Thus, while systematic sequential sounds form spoken languages of the world, structured systematic sequential images manifest in ***visual languages*** of the world.

These "visual languages" are thus used in the social objects which many cultures call "comics." However, visual languages also appear outside of this socio-cultural context, such as illustrated children's books, instruction manuals, and even aboriginal sand drawings. Comics stereotypically unite two separate languages to form a multimodal whole: a verbal/written language (produced via "writing") and a visual language (produced via "drawing"). However, because the patterns involved with visual languages may differ between the minds of populations of individuals across the world, no single visual language exists any more than a single spoken language. Rather, different visual languages are used in different contexts. For example, a particular dialect of American Visual Language characterizes the comics of mainstream superhero comics, which contrasts from that of the Japanese Visual Language stereotypically used in manga. This variation is evident both in their graphic structure (i.e., different "drawing styles") and underlying structures like panel framing and narrative patterns (Cohn, 2013b, 2015b) (Note: these visual languages *do not* define nor equate to comics or manga, they are simply characteristic of them). Because of this diversity, VLT posits that sequential images are not understood uniformly by all individuals. Rather, the understanding of sequential images depends on the *fluency* for particular visual languages—namely, those found in the specific comics that a person reads (Cohn and Kutas, 2017).

Within VLT, sequential images are posited as being understood via a ***Visual Narrative Grammar*** (VNG). This structure is separate from meaning, and functions to package semantics into well-formed as opposed to ill-formed sequences. That is, it constrains sequences to present meaning in ways that make sense, as opposed to those that do not. Like syntactic structure in sentences, this "narrative grammar" gives categorical roles to individual image units, and then organizes those units into hierarchic constituents (Cohn, 2013c). However, because images contain more semantic information than individual words, this narrative grammar operates at a "discourse" level of meaning, rather than at a sentence level. Nevertheless, the basic principles and cognitive mechanisms remain similar between the sentence (syntax) and narrative levels.

VNG uses several primary sequencing patterns (Table 1): *A canonical narrative schema, a conjunction schema,* and *a head-modifier schema*. In previous publications, these schemas were written as "rules" in the format of X → Y − Z (read as "X consists of Y and Z"). However, this is just a notational variation. As will be touched on, there is good reason to emphasize the schematic nature of these principles.

Table 1
Basic constructional patterns in Visual Narrative Grammar.

a) Canonical narrative schema:	$[_{\text{Phase X}}$ (Establisher) − (Initial) − Peak − (Release)]
b) Conjunction schema:	$[_{\text{Phase X}}$ X_1 - X_2 -… X_n]
c) Head-modifier schema:	$[_{\text{Phase X}}$ (Modifier) − X − (Modifier)]

VNG argues that aspects of meaning within images provide cues for narrative roles played by panels, and these categories organized into a canonical narrative schema. The basic narrative categories[2] include:

Establisher (E) − sets up an interaction without acting upon it, often as a passive state.

Initial (I) − initiates the tension of the narrative arc, prototypically a preparatory action and/or a source of a path.

Peak (P) − marks the height of narrative tension and point of maximal event structure, prototypically a completed action and/or goal of a path, but also often an interrupted action.

[1] For a counterpoint to B&W's claims regarding the combinatorial structure of individual images (such as "visual morphology" where elements like lightbulbs float above character's heads), see Cohn et al. (2016).

[2] Note: some categories such as Prolongations and Orienters are omitted for simplicity (Cohn, 2013c).

Release (R) — releases the tension of the interaction, prototypically the coda or aftermath of an action.

Though these narrative roles outline their prototypical meanings, full identification of a narrative category involves an interaction between a panel's content and its context in a global sequence (Cohn, 2013c, 2014c). This is similar to syntactic categories in language. Though syntactic categories (like nouns, verbs) have typical mappings to meaning (like objects, events), these are ultimately not how such categories are defined (Jackendoff, 1990). Rather, their definition relies on relational roles within the context of a sentence. Thus, the word "dance" (semantically, an event) can serve as either a noun (*the dance*) or a verb (*they dance*) depending on its surrounding context. Similarly, a panel of a passive state may act as an Establisher or a Release, depending on context (Cohn, 2013c, 2014c).

The basic narrative schema in VNG (Table 1a) organizes these narrative categories into a "phase" (a narrative constituent as a grouping of units) that uses a specific order. However, not all phases must contain all categories, and most elements are non-obligatory, as notated by the parentheses in the schema. Only Peaks are non-obligatory, because they motivate their local sequence (as its "head"), but Peaks can also be omitted under specific constrained, inference-generating contexts (Cohn and Kutas, 2015; Cohn and Wittenberg, 2015; Magliano et al., 2016; Magliano et al., 2015).

If only the canonical sequence order was possible, this would yield only short, simple narrative sequences. Thus, more complex sequences arise by narrative categories applying both to individual images and to whole groups of images. Consider the sequence in Fig. 1. This sequence starts with one boxer reaching back to punch another one. This preparatory action should cue this panel as an Initial, leading to the punch in the subsequent panel, a Peak. An Establisher then sets up a new action with the boxers facing each other passively. Another Initial shows a preparatory action, but in the penultimate panel the other boxer slips and falls, still a climactic Peak. The final Release panel shows the aftermath of this action, with the victor standing triumphantly over his fallen foe.

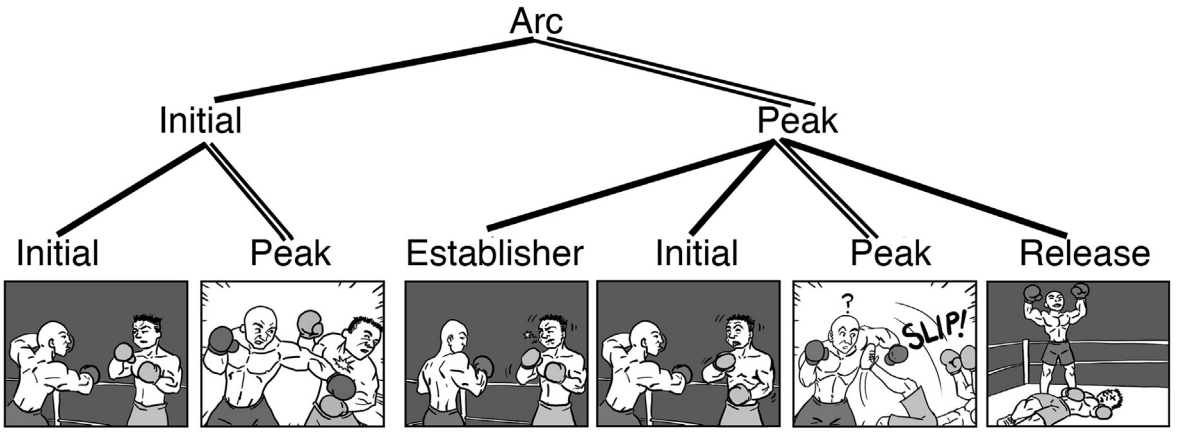

Fig. 1. A narrative sequence with two narrative constituents.

If we only consider the surface narrative categories of the panels (*I-P-E-I-P-R*), this sequence does not conform to the canonical narrative schema. Yet, by grouping panels together, these constituents can play relative narrative roles at a higher level. The maximal node is considered an "Arc," as it plays no role in a higher structure. The sequence in Fig. 1 has two constituents, an Initial of the first two panels, which sets up a Peak of the remaining four panels. Thus, each constituent internally follows the narrative schema, with Peaks motivating their local phase (i.e., the "heads" of the phase, indicated by double-barred lines). Thus, narrative roles apply to both individual panels and whole constituents. Embedding can also extend further upward to larger and larger levels of narrative, since these principles are recursive. Thus, the same constructs that might operate for a strip or a sequence within a comic can also account for higher "plot" level structures for a longer story.

While these notions are at least somewhat similar to traditional notions of narrative—albeit more operationalized—other schemas in VNG introduce further complexity to sequences. For example, ***conjunction*** (Table 1b) repeats narrative categories within a constituent of the same category (Cohn, 2013c, 2015b), analogous to syntactic conjunction which repeats grammatical categories (like multiple nouns in a noun phrase: *The fork, knife, and spoon*). Fig. 2b uses conjunction in its first three panels, which show three characters, each in their own panel. These three panels in (2b) depict the same conceptual information as the single first panel in (2a), suggesting that they all play the same functional role (as an Establisher, introducing the scene). Indeed, these three panels in (2b) could be substituted for the one in (2a). However, though their individuation draws focus to each character, no overt cues indicate that they belong to the same spatial location (as in the panel in Fig. 2a), and thus we must infer this shared spatial environment. Such a process of inference occurs by updating a mental model of the scene, which interfaces with the narrative pattern (Cohn and Kutas, 2017). This type construction is called ***Environmental-Conjunction*** or ***E-Conjunction*** since the conjunction (repetition of panels playing the same narrative role) maps to an inference of a broader spatial environment (notated with subscript "e").

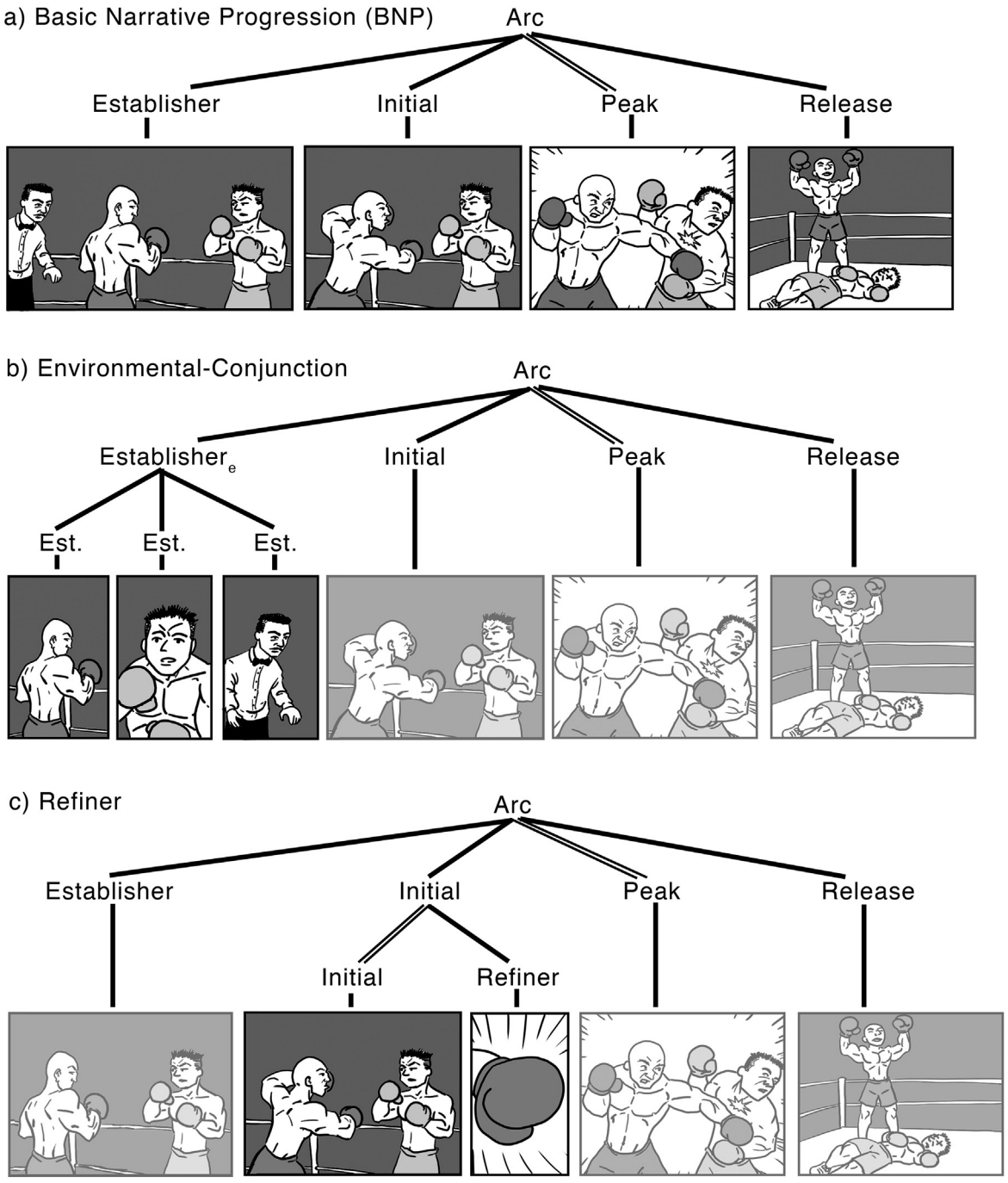

Fig. 2. Different basic narrative schemas within Visual Narrative Grammar.

Conjunction can also convey other types of information beyond scenes. Narratively, conjunction only specifies that categories repeat within a constituent. Thus, this *narrative* pattern (repeated panel category) allows for various mappings to *semantics* (Cohn, 2013c, 2015b), several of which are depicted in Fig. 3. Each three-panel conjunction on the left tier can serve as Initials in the sequence. These panels can show (a) actions or events (Action/A-Conjunction), (b) characters within a scene (Environmental/E-Conjunction), (c) parts of a single character (Entity/N-Conjunction), or (d) disparate semantically associated elements (Semantic Network/S-Conjunction). Such semantic information is connected to the conjunction schema using correspondence rules that specify both panel content and the superordinate information that would be inferred as part of constructing a mental model of the scene (Cohn, 2015b). These conjunction sequences (left tier) are meaningfully equivalent to the single panels in the right tier, which can also serve as Initials in this structure. Thus, a creator can choose from all the

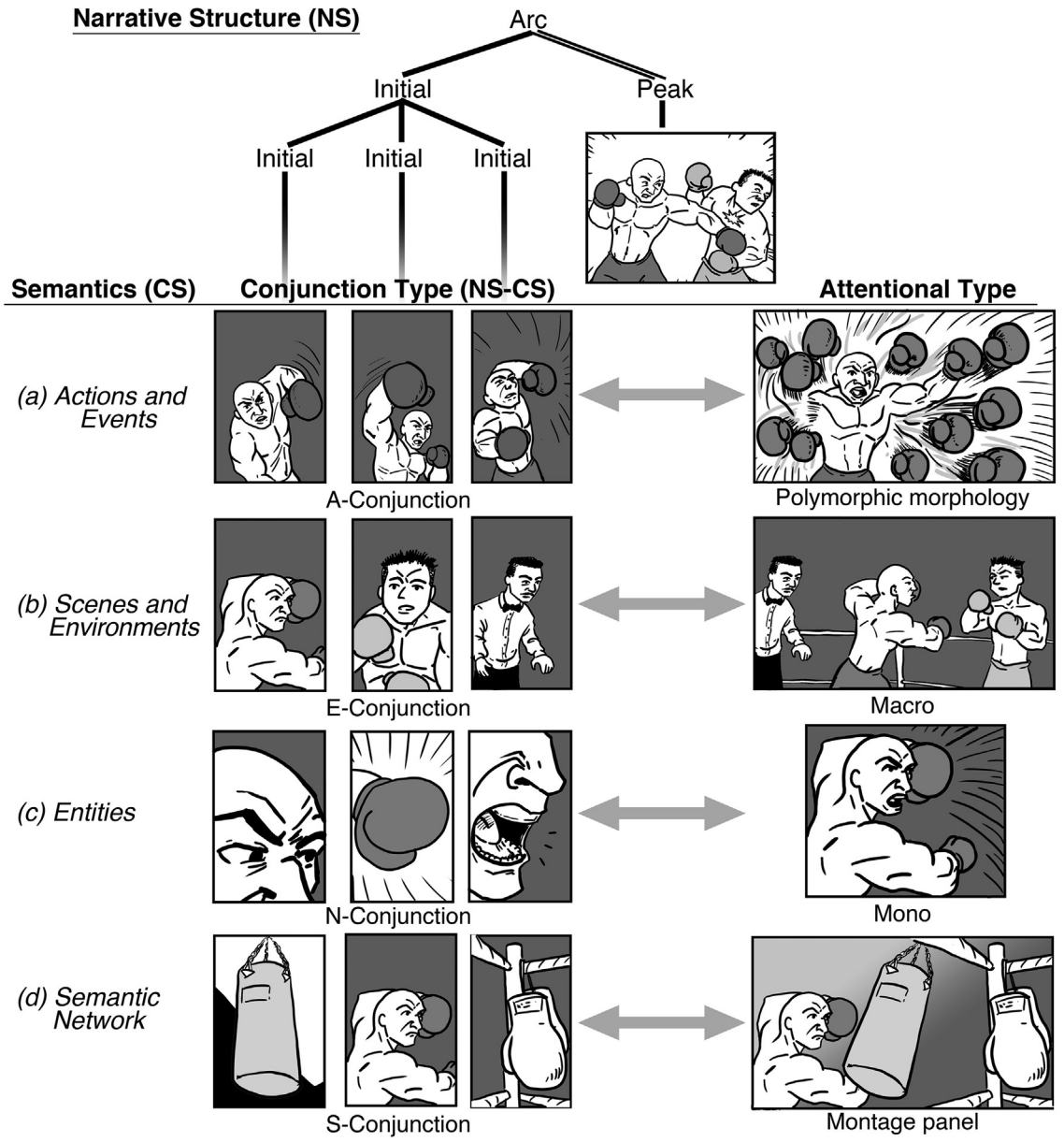

Fig. 3. Various semantic options for a narrative Initial, either as a conjunction sequence (left tier) or a single image (right tier) (Cohn, 2015b).

options in the left tier or the right tier to fill that narrative "slot" of an Initial. More options come from mixing and matching these structures. For example, you could replace the first panel in 3b (E-Conjunction) with all three-panels in Fig. 3c (N-Conjunction), to create a conjunction embedded within a conjunction.

It is also worth noting that the two tiers in Fig. 3 provide tests for each other. Because the right and left tiers are posited as equivalent, those single images should be able to substitute for their corresponding conjunctions. In analysis, this serves as a diagnostic: You can test if a sequence uses an E-Conjunction, for example, if it semantically can be replaced by a "macro" panel that contains the same information (Cohn, 2015b). Such a "substitution test" is one amongst many diagnostics used to confirm the structure of a sequence, in line with the logic of diagnostic tests of syntax (Cheng and Corver, 2013), which provide the basis for psychological experimentation (Cohn, 2013c, 2014a, 2014b, 2015a).

Another modifier comes from panels which can draw focus to information in other panels. The third panel in Fig. 2c zooms in on information in the preceding panel (the puncher's fist), a ***Refiner*** (Cohn, 2013b, 2015b). Refiners modify the information in another "head" panel (again, double bar lines), and therefore, unlike conjunction, Refiners do not repeat the same narrative role as their head. Rather, they modify the head panel with added focus, while the head retains its wider viewpoint and more fundamental role in the sequence (as in 3a). As outlined in the head-modifier schema (Table 1c), Refiners can go either before or after their head.

The combination of these schemas allows for VNG to account for a wide variety of complexity within visual narratives. Consider Fig. 4a, from *Vagabond #15* by Inoue Takehiko. First, note that the layout has been altered to be linear—this is done only for ease of notating the narrative structure. The structures that operate on page layout are separate from narrative/meaning (cf. B&W), as evident in experiments showing that participants have preferences for navigating layouts even in the absence of content (Cohn, 2013a; Cohn and Campbell, 2015). In addition, altering the same content into different page layouts has minimal influence on sequence comprehension (Foulsham et al., 2016; Omori et al., 2004). Thus, the structures governing page layout are interfaced with those of narrative, but not dependent on each other (Cohn, 2014a). Such layout structures are omitted here for simplicity.

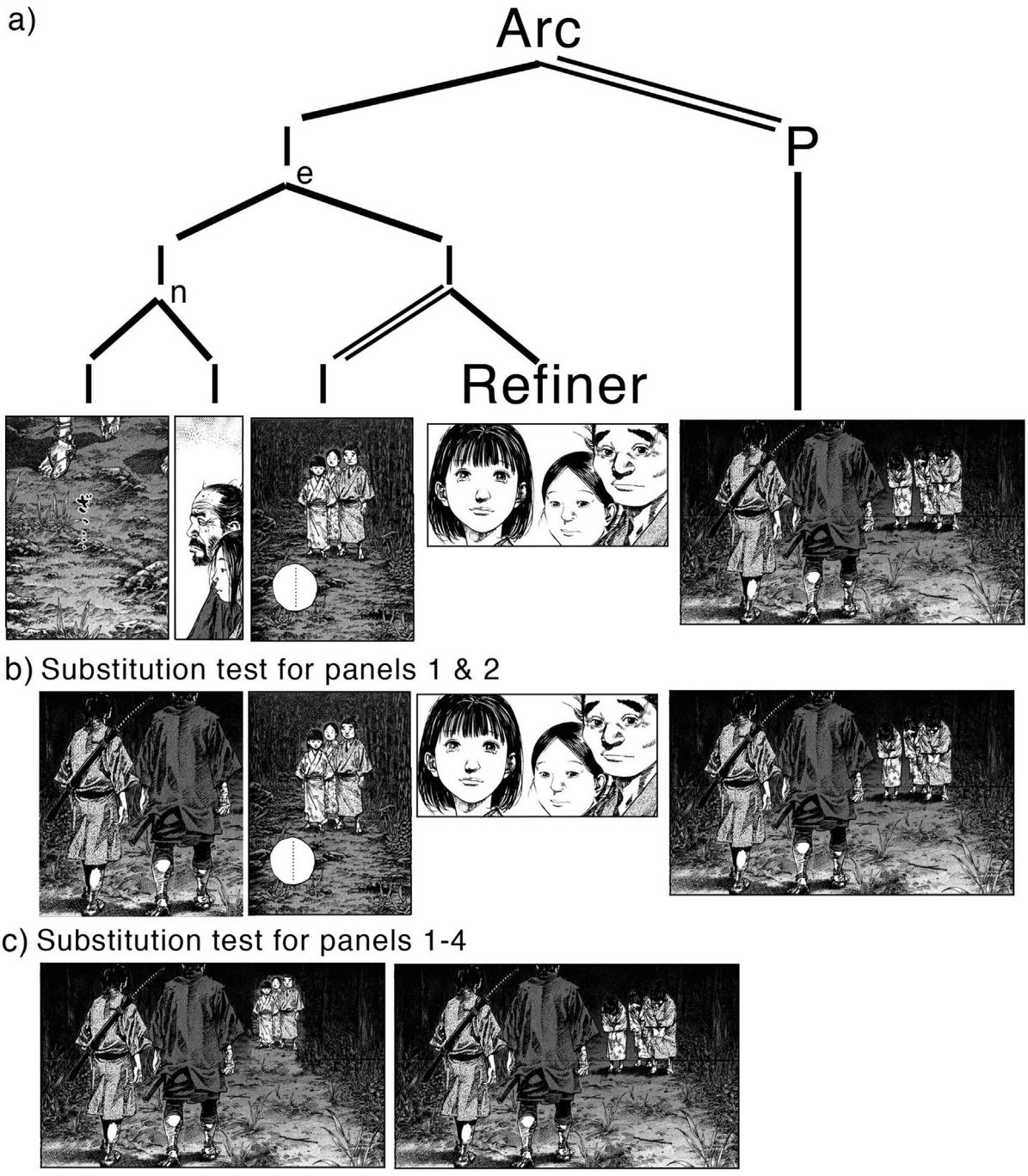

Fig. 4. Narrative sequence from *Vagabond* by Inoue Takehiko (Vol. 15, Kodansha), which uses N-Conjunction, E-Conjunction, and a Refiner, all within the canonical narrative arc.

The content of the sequence in Fig. 4a conveys a fairly simple action. A man and boy walk down a road and meet a family, who bows to them (in thanks for saving the girl's life, in an earlier scene). The first two panels show only glimpses of the man and boy's body, whether their feet (panel 1) or busts (panel 2). These panels are combined using N-Conjunction, to construct the notion of the whole entities (notated with subscript "n" on the constituent). As described above, N-Conjunction can be tested by replacing these two panels with a single panel of just those characters, as in Fig. 4b.

The subsequent panels focus on the family, first in a whole frame (panel 3) and then zooming on their faces (panel 4). This penultimate panel is a Refiner, which modifies the previous panel by repeating a portion of the same information at a closer view. Yet, the man and boy are never shown in the same space as the family in these four panels. Thus, each constituent is then united together using E-Conjunction to form a larger grouping (subscript "e"). Again, E-Conjunction can be tested by replacing those panels with a single panel of equivalent content, as in Fig. 4c. The final panel shows the family's action of bowing. All the preceding panels set up this Peak panel, as narrative Initials. Again, the interpretation that these are all Initials is provided by the substitution test in Fig. 4c: all can be replaced by a single panel. Note also that the canonical narrative arc is maintained in the upper level, though not all categories are used (there is no Establisher or Release here).

Thus, this example can hopefully show that VNG can effectively characterize various structures in actual sequences from comics, even those that are fairly complex. VNG does this using basic schematic patterns: the canonical narrative schema, the conjunction schema, and the head-modifier schema. These patterns are similar to those posited by theories of syntax at the sentence level (Culicover and Jackendoff, 2005). These sequencing patterns thus map to different meanings (and at a different level of semantics than sentence-level syntax), to allow for a variety of surface manifestations. These constructs can be tested using diagnostics, such as a substitution test, which are common in linguistics research (e.g., Cheng and Corver, 2013). In addition, because VNG is a "construction grammar" (discussed below), it also allows for patterned sequences that do not necessarily conform to these basic schemas. If a sequencing pattern is regularized and systematic, it also may be stored in creators' (and readers') memory as part of this visual language grammar.

3. Grammars

We now turn to the critique of VNG made by B&W and their own model. B&W's primary criticism lies in the overarching orientation of VNG as being a "grammar." They state even in the abstract that "the notion of 'grammar'…is in many respects deeply problematic" (Bateman and Wildfeuer, 2014a, p. 373) and subsequently that "Very few of the properties exhibited by grammars of natural language in fact carry over to visual media such as comics – the problem of using the 'grammar' metaphor is then that it leaves unclear just where the boundaries of the application of the metaphor should be placed" (p. 374).

Despite this being their primary criticism, nowhere in either publication do B&W articulate just what they mean by "grammar", its basic assumptions, or what is objectionable about it. At most, they describe grammars to be undesirable because they impose "rigidity" to analysis because of "concerns over the appropriateness" of "functionally distinguished structural slots" (Bateman and Wildfeuer, 2014b, p. 200). Yet, in no place do they actually discuss the "properties exhibited by grammars" nor that "visual media such as comics" lack them.[3] Rather, "grammar" is stated an undesired construct, without directly engaging why (though they allude to a previous generalized literature, which may or may not be applicable to VNG). As will be discussed further on, this renders "grammar" as a strawman argument by which they can level critiques in favor of their own approach.

Clearly defining what is meant by "grammar" is important in many respects, not the least of which being that linguistics itself has different grammatical models. There are several models of grammar in the study of syntax, and these models carry different assumptions and principles. B&W seem to implicitly assume that VNG is "generative grammar" in the tradition of Chomsky (1965, 1981). This type of grammar holds that a memorized list of vocabulary items (the lexicon) combines with algorithmic rules for ordering (phrase structure rules), which then outputs a basic syntactic structure. From this, semantics is interpreted based on the syntactic structures. Here, a sequence is deemed "ungrammatical" if the algorithm fails to properly generate the appropriate syntactic structure.

The belief that VNG is a Chomskyan grammar is perhaps not unfounded. Chomskyan theories have been dominant and widely known from linguistics over the past several decades, and to many the sheer presence of "tree structures" as in Fig. 1 are taken to indicate such a model. For those less familiar with syntactic models and debates about them, it is understandable for any grammatical model to be mistaken as reflecting the Chomskyan approach, given its renown. In addition, precedents have proposed formal "grammatical" approaches for verbal narrative stories (e.g., Mandler and Johnson, 1977) and for film (e.g., Carroll, 1980) based explicitly on Chomskyan models.

However, VNG is *not* a Chomskyan grammar. Rather, VNG is based on contemporary theories of *construction grammar* (Culicover and Jackendoff, 2005; Goldberg, 1995), and is indeed integrated into these models explicitly (Cohn, 2016). Construction grammars differ from Chomskyan models in several ways. First, in the particular model of construction grammar followed by VNG—i.e., Simpler Syntax (Culicover and Jackendoff, 2005)—grammar and meaning are independent, yet mutually interfacing components. Thus, meaning is not interpreted from grammar, but is a separate structure that

[3] Note also: in VLT, it is a misnomer to say that "comics has a grammar." Only visual languages have grammars, and such visual languages appear in comics. The analogue would be saying that "novels have grammar" rather than that written language has a grammar and is used in novels.

coalesces with grammar (or specifically, is *organized by* grammar). This separation is evident in the various meanings (actions, environments, entities, semantic fields) that map to the singular grammatical conjunction schema, as illustrated in Fig. 3.

Second, sequencing does not arise out of inserting memorized lexical items (like words) into phrase structure rules. Rather, in construction grammar, "rules" are stored in long-term memory as schemas unto themselves (Goldberg, 1995; Jackendoff, 2002). In other words, they are not so much "rules" as they are cognitive patterns, just as words are a type of cognitive pattern. Thus, "ungrammaticality" (or "well-formedness") here is a case of how well a sequence conforms to the expectations of a memorized pattern, not an error in computation. In addition, while abstract grammatical patterns do exist (like those in Table 1), construction grammars also allow for unique sequencing. For example, English uses a construction of *VERB-ing the TIME away* in phrases like *twistin' the night away* or *reading comics the evening away*. This is a memorized pattern with empty "slots" that does not necessarily conform to abstract sentence structures. Similarly, as mentioned above, unique narrative constructions may also exist for both images and multimodal constructions, such as sequence patterns using a silent penultimate panel followed by a punchline (Cohn, 2013b).

Because VNG is a construction grammar, it carries different assumptions and expectations about its structure than if it were a Chomskyan phrase structure grammar. B&W do acknowledge that VNG follows the model of Jackendoff (2002) but it is unclear if they recognize that it thereby uses a construction grammar (Bateman and Wildfeuer, 2014b, p. 199), and whether it would change their critiques. As it stands in their criticism, "grammar" provides the spectre of something theoretically undesired, albeit without ever explaining why or showing evidence of its inadequacies. This rhetorically serves to justify the (unstated) benefits of their own approach, to which we now turn.

4. Bateman and Wildfeuer's discourse model

As an alternative to VNG, B&W offer a model based instead on Segmented Discourse Representation Theory (Asher and Lascarides, 2003). Here, meaningful relations between images are characterized by properties such as *Narration*, *Parallelism*, *Detail*, or *Contrast*. Such relations are similar to the "panel transitions" posited by theorists like Scott McCloud (1993) and other approaches to visual narrative based on discourse coherence models (Saraceni, 2000, 2016; Stainbrook, 2003, 2016). However, unlike these precedents, which maintain meaningful relations only by *linear juxtaposed* units, B&W also posit that such relations can operate hierarchically. This allows them to claim that their model generates the same hierarchic relations as VNG, only it does so without the need for a grammar. Furthermore, they also claim that this model can account for non-linear relationships, and thereby can characterize image relations across a whole page layout.

The B&W approach is based around "defeasible hypotheses" for interpreting the meaningful relationships between juxtaposed images. Each image is compared to a subsequent image, and their possible meaningful connections are noted via specified restrictions of relational content. For example, Fig. 5 shows sequences taken from Bateman and Wildfeuer (2014a), modified from those in Cohn (2010). In panel 1, a man reaches back to punch, followed by panel 2 of another of a man looking nervous, and then panel 3, which zooms in on the eye of the man in panel 2. In B&W's model, the first two panels show a *Contrast*, because they change between "semantically different" elements of the first and second characters (Bateman and Wildfeuer, 2014a, p. 389). The third panel shows a close-up of the second character, and thus panels 2 and 3 connect using a *Detail* relation, a type of *Elaboration* (Note that B&W's publications differ on this function: Bateman and Wildfeuer (2014b) calls it "Part of" while Bateman and Wildfeuer (2014a) calls it "Detail," an irregularity that is not explained.)

Fig. 5b expands this sequence with additional panels: another zoom (now of the man in panel 1) appears as panel 3, pushing the original zoom to become panel 4. In addition, a final panel shows the first man punching the second. Now, the first two panels again show a *Contrast* between characters, but so do the zoomed images in panels 3 and 4. Each of these pairs then forms a grouping, which are linked to each other through a relation of being *Parallel*. This relation specifies that each group maintains the same order of referenced characters (i.e., an A-B-A-B order). This link thus creates a larger grouping comprised of all four panels, which stands in relation to the final panel through *Narration*, since the sequence then progresses in time.

Thus, unlike other characterizations of image-to-image panel transitions (e.g., McCloud, 1993), B&W's approach allows for semantic relationships to form hierarchic clusters. However, B&W emphasize that such organizations are not codified structures. Rather, they are meant to characterize the "defeasible hypotheses" of a reader's interpretation. Other interpretations are conceivable if a reader views these relations in different ways. In addition, as emphasized by the left-to-right arrows, this approach is meant to characterize the "dynamic" nature of interpretations as it unfolds panel by panel in the process of reading.

5. Comparing approaches

There are several ways in which B&W's approach does not adequately critique VNG, and is insufficient as an alternative model. How might we adjudicate the value of a theory? One of the arguments in B&W is that their model's capacity to generate comparable structures to VNG without appealing to a grammar. This appeal to *parsimony* has a long tradition in linguistics theory (e.g., Chomsky, 1972; Postal, 1972), and assessing the relative value of a theory with or without a grammar would echo such debates. However, efficacy of a theory can be judged across several dimensions, parsimony being only one of them (Culicover and Jackendoff, 2005, p. 4; Ludlow, 2011, p. Chapter 7), and before addressing theoretical elegance, it is first

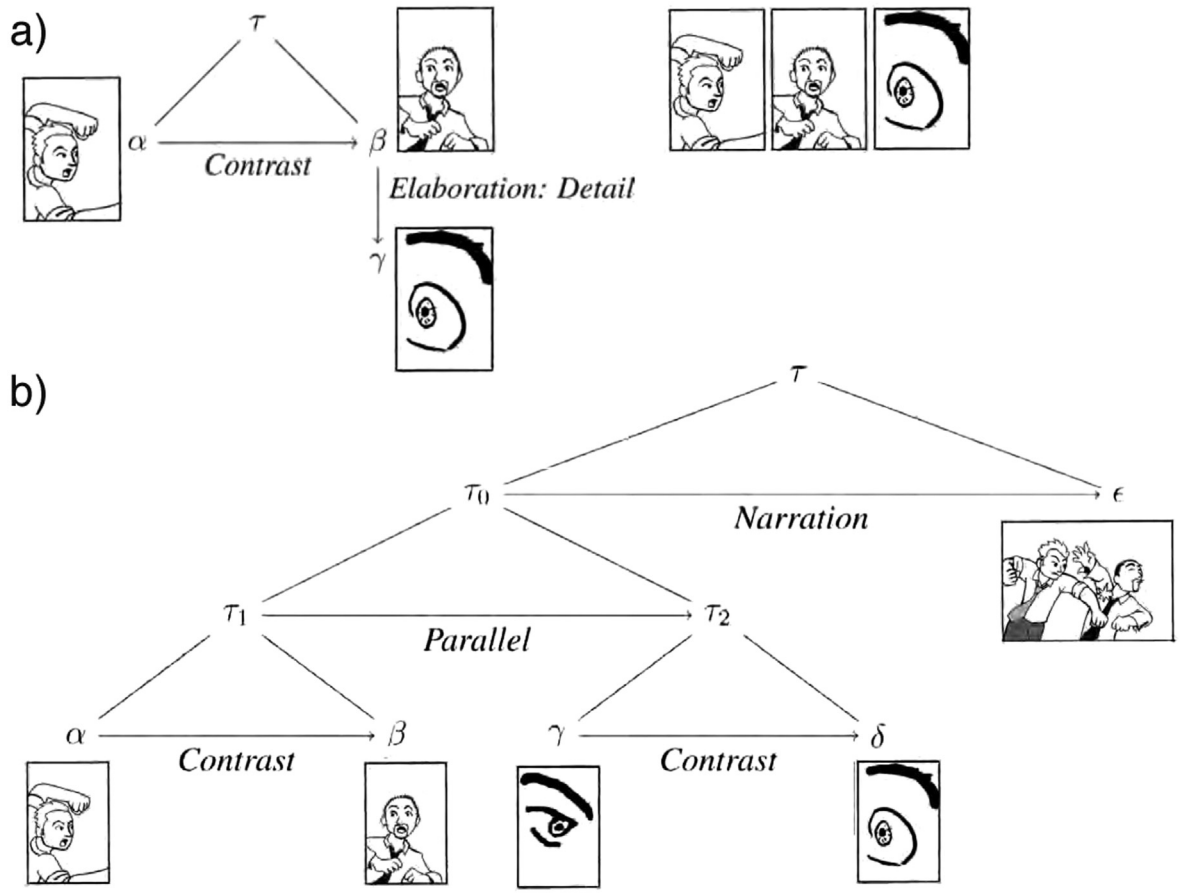

Fig. 5. Dependency trees characterizing relations between panels in Bateman and Wildfeuer (2014a). a) Example where a zoom panel is treated as an elaboration. B) Example where zoom panels no longer retain their properties as elaborative details (figure modified with enlarged images).

important to ask how well a theory accounts for the data. First, a theory should be able to sufficiently describe the phenomena that exist within a system. In this case, can a theory effectively describe what occurs in the visual sequences of comics? Second, a theory should be testable using empirical methods, such as psychological experimentation. If a theory is meant to characterize actual cognition, experimental evidence provides the way to test it, and a theory should fit such experimental observations.

Below, we review the ability of B&W's model to adhere to these criteria in contrast to the assumptions made in VNG. Note, in their papers B&W never make such a direct comparison between their approach and VNG, though they state that, "A more exact comparison of the two approaches would therefore be very worthwhile" and then hedge that despite potential for utility in the VNG approach, "further empirical investigations…are clearly necessary" (Bateman and Wildfeuer, 2014b, p. 200). Their own discussion of the differences between approaches remain in abstract terms (e.g., a discourse approach is better because it's "not grammatical"), and often characterize VNG by principles in outdated, earlier models (Cohn, 2003, 2010), with minimal discussion of how those theoretical differences change their interpretation. This lack of comparison means that they never show cases where their model makes observations that are absent in VNG, or how it better supports—or even addresses *any* of—the existing empirical data. This is also important because B&W claim that their approach can capture the same basic observations as VNG, yet they never directly address the constructs in both models. Such a comparison is thus made below.

5.1. Descriptive adequacy: Comparison #1

Consider again Fig. 5b. This dependency structure yields an *A-B-A-B* pattern, where, similar to VNG, each "A-B" pair is grouped together (panels 1/2 and 3/4). In VNG, this would be captured by successive E-Conjunctions (Fig. 6a). In B&W, it is made by noting the semantic *Contrast* between characters, with these constituents linked through a *Parallel* relation. However, the previous relation observed in Fig. 5a of *Detail* has now disappeared. Even though panels 2 and 4 are the same in 5a and 5b—with ostensibly the same relationship—the zooms are no longer observed as displaying *Details* of their antecedents because a single panel distance separates them. That is, B&W's model does not capture that 3 is a detailed zoom of 1 and 4 of 2. Is this distance relation no longer interpreted as such when a sequence uses two zooms instead of one? By appearances, the B&W model seems to fail at consistently describing the properties even in their provided examples.

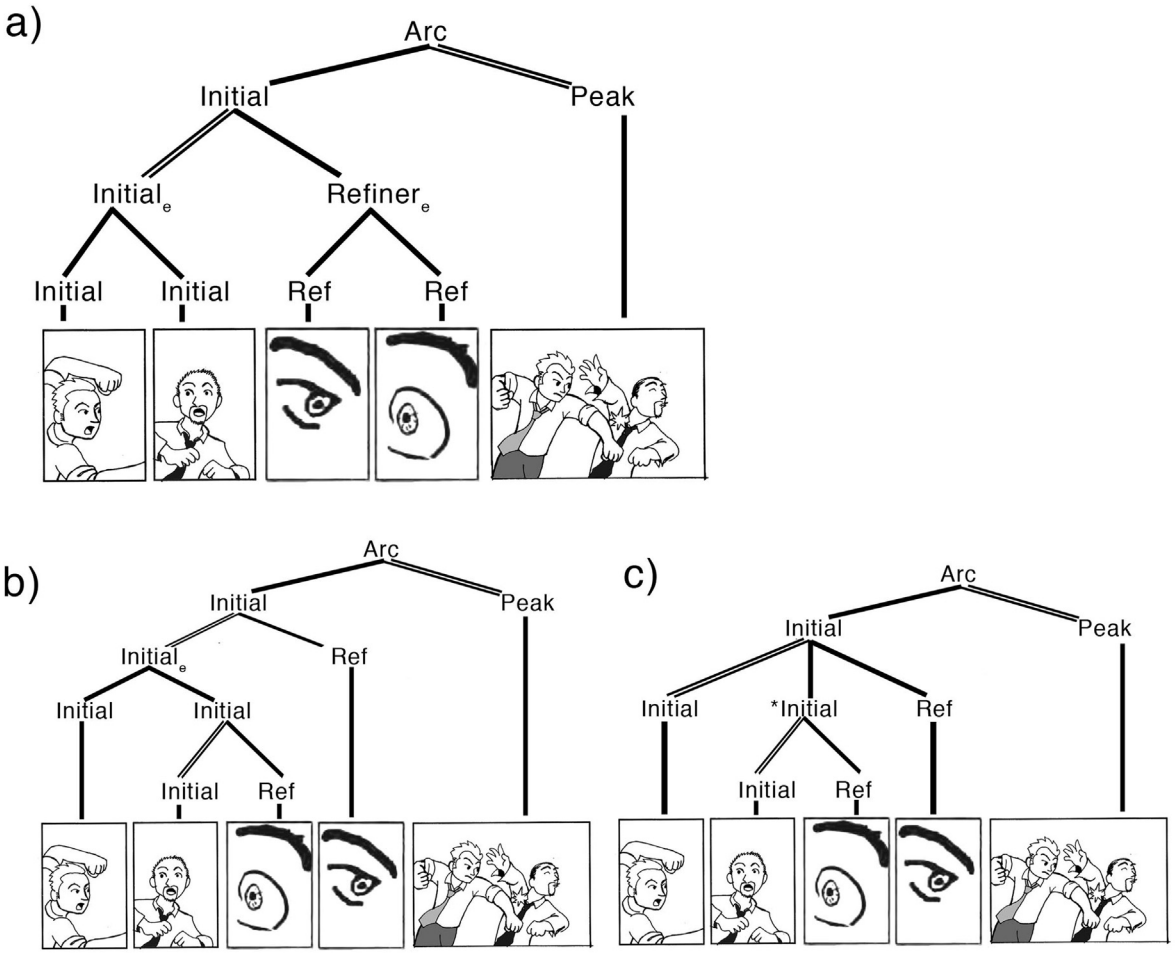

Fig. 6. Sequence patterns in VNG using a) ABAB patterns or b/c) ABBA patterns.

Let's compare how VNG handles this sequence. As in Fig. 6a, VNG notates each pairing of panels using E-Conjunction (B&W's observation of *Contrast*). The first pair play the role of Initials (in relation to the final panel Peak), while the second pair are Refiners. Again, conjoined relations could be tested using a substitution test, as described above. In relation to the Initials, the Refiners hold the relationship of "zooms" (B&W: *Detail/Part-of*), which is the absent relation in B&W discussed above. Refiners could be tested using an "inset test" (Cohn, 2015b): Can the same content be conveyed by drawing an inset panel in their heads? The superordinate constituent here notes that these Refiners are in relation to the preceding Initials, with these specific relationships percolating into the subordinate constituents.

It is worth noting that, as B&W observe, the basic hierarchic relations are the same in VNG and B&W's discourse approach. However, VNG also captures observations about the narrative roles that images play relative to each other: the Refiners play roles *as modifiers of* Initials, and the Initials play roles *as expansions of* the Peak. In VNG, the similarity between roles of images can be tested by diagnostic tests like substitutions (Cohn, 2013c, 2014a, 2014b, 2015a), where images of like-categories can substitute for each other. These relations are the reverse of those in B&W. In VNG, panels play roles, and those roles are linked through positions defined by a structural schema, but the relations themselves are not characterized (there is no "Initial-to-Refiner" relation or "Refinement" relation). In B&W, images play no roles, and relations themselves are given substantial identities. This difference is taken up in more detail below.

This structure in VNG also captures the primacy of some panels compared to others—"heads" are more central to the sequence, which is why the second set of panels (Refiners) can be omitted more felicitously than the first pair (head Initials). If the first pair is omitted, the Refiners take the roles of heads (Initials) and may simply be less explicit about their content. This absorption has been structurally compared to adjectives (a modifier) acting as nouns (a head) as in *I'll have the red* to mean *red wine* (Cohn, 2013b, p. 85). B&W's approach would not explain why some images can delete with more or less well-formedness—any manipulation should yield an equally defeasible interpretation as any other. Thus, though the surface "similarity" in hierarchic structure may exist, VNG captures additional insights beyond the B&W approach.

One observation appears in the B&W approach that does not arise overtly in VNG though: that the successive pairs of panels are "parallel" to each other. That is, they both use "AB" patterns in succession. In VNG, the superordinate structure illustrates the Initial-Refiner head-modifier structure, but does not dictate the relations within its subordinated constituents. Thus, on the surface, VNG does not make this observation in the same way that B&W do not observe the distance *Detail/Part-of* relations.

Given this, consider the sequence in Fig. 7, which adds a panel with a third character after the first two panels to the sequence from Figs. 5b and 6a. Now there are three characters (panels 1–3) followed by two zooms (panels 4 and 5). In VNG, this manipulation simply adds an Initial panel to the E-Conjunction, while nothing else changes in the structure (Fig. 7a): the Refiners still maintain the same relation to the prior Initial phase as before. In the B&W approach, this would greatly change the structure (Fig. 7b). Here, the additional character (notated as "X") would presumably create another *Contrast* locally, but this addition would make the *Parallel* relation disappear, since the two groupings no longer maintain the same order of similar information. In Fig. 7b, this relation is left as a question mark so as not to impose an interpretation within the B&W approach (though the groupings are maintained, which they may also deem needed to be altered). However, when B&W are similarly faced with such an alteration by reversing the position of the zooming panels to create an ABBA pattern (as in Figs. 6b and 8c), they also do not state the resulting relation. Rather, they posit multiple possible interpretations because such an alteration would need "more contextual information to restrict the range of potentially viable interpretations further" (Bateman and Wildfeuer, 2014a, p. 395). It is unclear why one sequence requires more context for construal, but another—containing all the same information—does not.

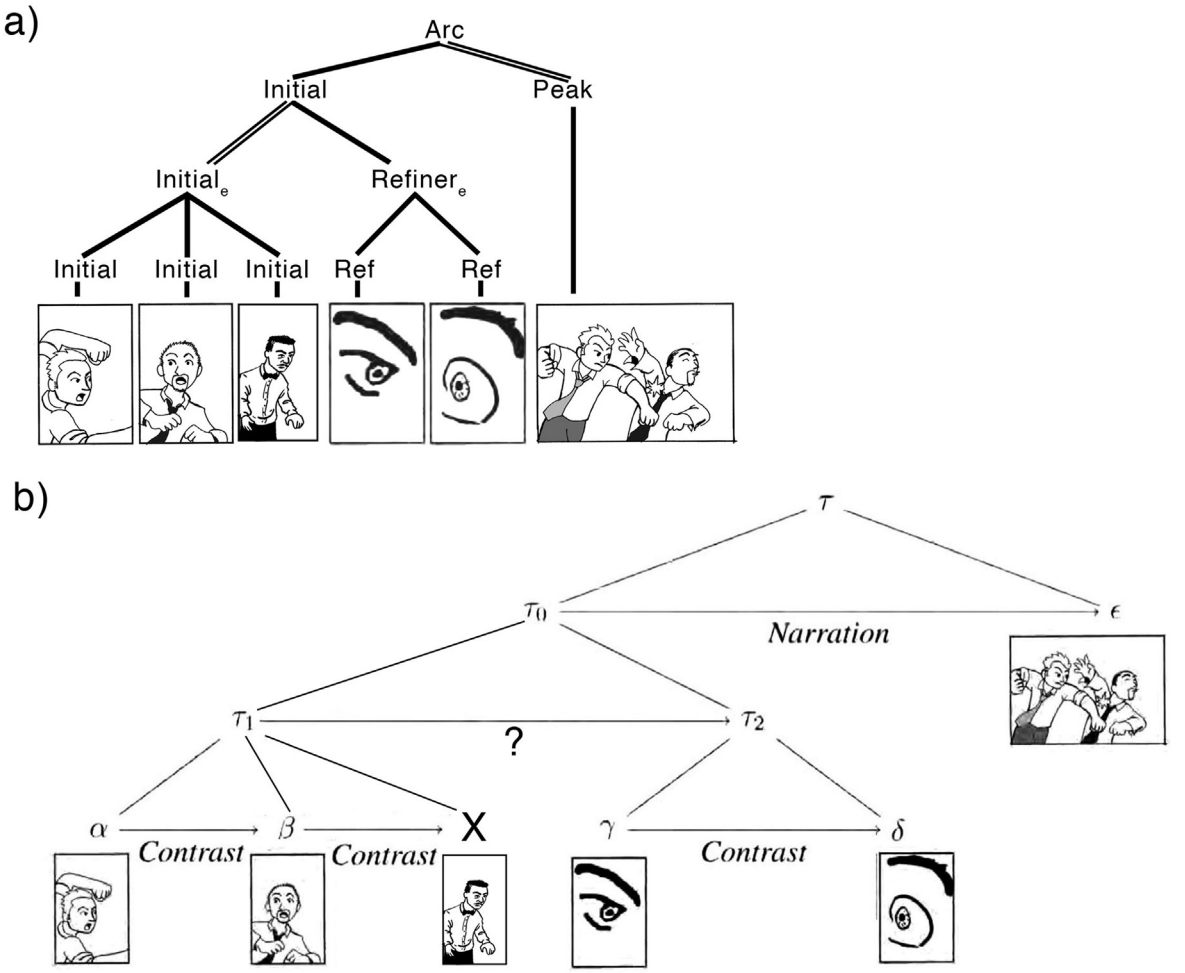

Fig. 7. Comparison of approaches for adding a character to the sequence in Figs. 5 and 6.

In VNG, the issue of parallelism is more complicated. The "parallel" constraint emerges from a mapping of narrative grammar to semantics (not depicted here, though see Cohn, 2015b), rather than a defined relation in the structure. Indeed, the content (i.e., the characters) is a facet of semantics, though the grammar links to these meanings. A constraint for alternating

structures thus emerges in the interface between grammar and meaning for an important reason. While the successive E-Conjunction structure in Fig. 6a could persist for an ABBA sequence, without this ABAB parallelism, adjacent "B" panels can "attract" each other to create an "island." In other words, the "BB" panels form a grouping that separates the A panels. This can create a parsing whereby one head-Refiner grouping embeds within another. This either yields a structure where Refiners and E-Conjunction alternate to create many successive embeddings (Fig. 6b), or form an ill-formed structure (6c). That is, an ABBA structure creates challenges for parsing the sequence.

A similar phenomenon happens with successive conjunctions in syntax (a *serial order dependency*). For example, the ABAB structure of *The yellow and red bananas and apples* should thus be easier to comprehended than *The yellow and red apples and bananas*. In this latter ABBA sequence, the juxtaposition of *red* and *apples* pressures the system to group them together because of their associated semantics, which would create an island between the "A" units: [A[BB]A] instead of [AB][AB]. This is directly analogous to VNG, and indeed the same constraint would be argued to occur in both. "Parallel" structures arise to prevent islands, as a byproduct of an optimal processing strategy. Note, this observed alignment between the structure of narrative and syntax contrasts B&W's unsubstantiated claim that "very few of the properties exhibited by grammars of natural language in fact carry over to visual media such as comics" (Bateman and Wildfeuer, 2014a, p. 374).

If such constraints do operate on visual sequences, then we would expect that ABBA structures would be worse than ABAB structures—as in Fig. 6. In fact, B&W discuss this type of pattern, as depicted Fig. 8c (their Fig. 8iii), as well as an AABB sequence in Fig. 8b. Though they describe several "defeasible hypotheses" for these ABBA and AABB sequences—and make a plea for needing more context for a "potentially viable interpretation" (Bateman and Wildfeuer, 2014a, p. 395, p. 395)—these sequences are not necessarily described as less preferred than the original ABAB sequence (Bateman and Wildfeuer, 2014a, p. 391). As stated above, VNG explicitly predicts ABAB sequences are more well-formed than ABBA sequences.

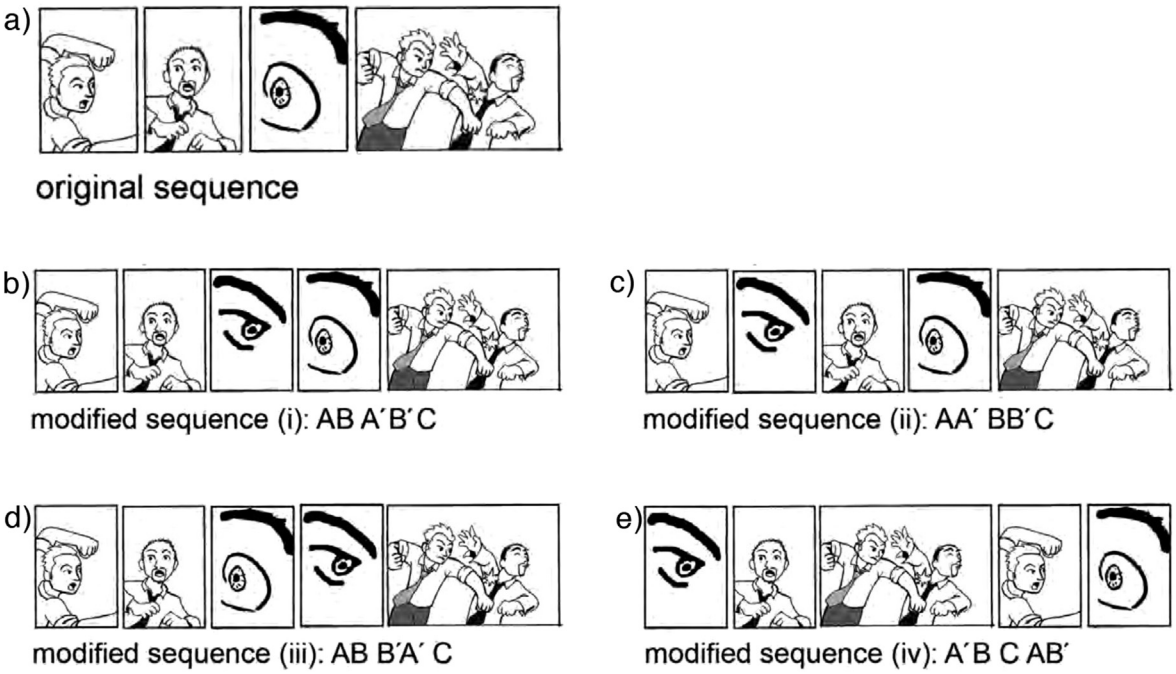

Fig. 8. Example manipulations from Bateman and Wildfeuer (2014a, Fig. 8). Letters added.

The felicity of ABBA sequences was recently tested in a pilot study using a "forced choice" test where participants (N = 33) were asked to choose which of two types of sequences they preferred. If both sequences are equally construable—as implied by B&W (Bateman and Wildfeuer, 2014a, p. 391)—then no preference should be seen for ABAB sequences over other patterns (i.e., each should be preferred at a rate of chance: 50%). However, we found that 78.5% of participants preferred ABAB sequences (like Fig. 8a) to AABB ones (like Fig. 8b), and 90.5% preferred ABAB sequences to ABBA ones (like Fig. 8c). Such preliminary results thus support that 1) some sequencing *of the same panels* are more preferred than others, and 2) that the ABAB sequence specifically is preferred, as predicted by VNG.

5.2. Descriptive adequacy: Comparison #2

Let's consider an additional example where B&W and VNG would diverge in their analyses. Fig. 9 has four sequences, which appeared as stimuli in Cohn and Kutas (2015). Here, Charlie Brown throws a ball (panel 1) which Snoopy then runs to

chase (panel 2). The critical third panel differs between strips, and each one changes the interpretation at the final panel. In both Fig. 9a and b, the panel depicts only Charlie, so that Linus's appearance in the last panel is unexpected. In both strips, an inference is generated that Linus retrieved the ball, not Snoopy, prior to that final panel. Fig. 9b suggests this event occurs concurrently, off-panel to panel 3, given Charlie's expression and the bubble (!), while Fig. 9a makes no hint of this character switch. Fig. 9c shifts perspective to show Linus at this critical panel, making his appearance in the final panel fairly unrevealing. Finally, Fig. 9d shows the expected outcome of Snoopy retrieving the ball at panel 3, only to switch at the final panel incongruously.

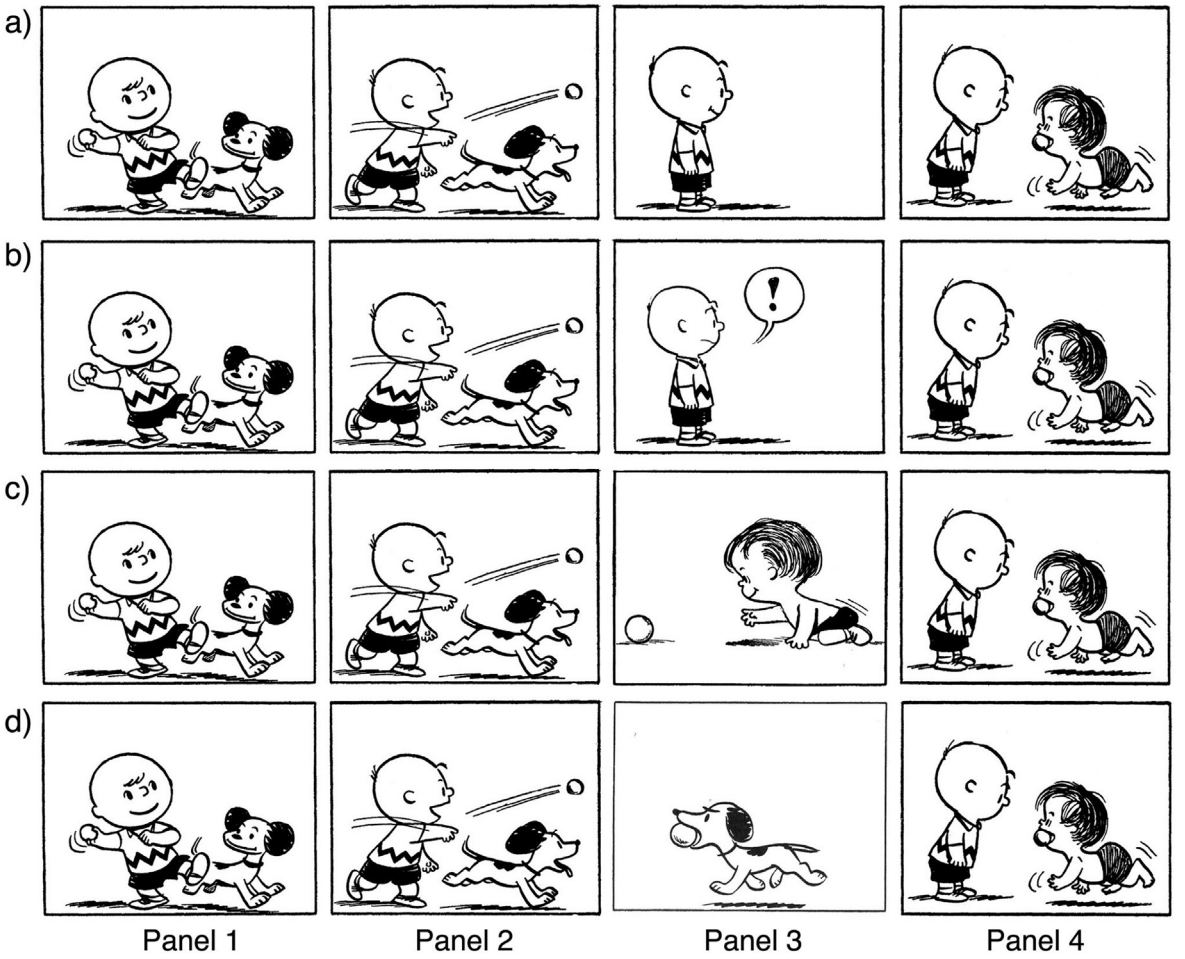

Fig. 9. Sequences used as stimuli in Cohn and Kutas (2015). *Peanuts* is © Peanuts Worldwide LLC.

These sequences should warrant different interpretations by the two models, crucially related to the status of panels 3 and 4. In B&W's approach, *Narration* relations should manifest between panels 2 and 3 for Fig. 9a, b, and d, since all use a clear temporal progression. Fig. 9c should differ, instead using a *Contrast* relation, by switching characters to Linus. *Narration* should subsequently arise between panels 3 and 4 for Fig. 9a, b, and c, clearly showing temporal changes, with *Contrast* in Fig. 9d switching from Snoopy to Charlie and Linus.

VNG contrasts sequences differently. In all sequences, VNG would categorize the final panel as a Release—where the final "tension" of the strip dissipates, and where the punchline occurs. In Fig. 9b, c, and d, panel 3 should be categorized as a Peak, since it depicts the primary events of the sequence. However, Fig. 9a would be underspecified, and would canonically be classified as a Prolongation, a medial category that functions to extend the payoff before a Peak panel (Cohn, 2013c, 2014c). Thus, in Fig. 9a, the Peak is fully elided, while in Fig. 9b, Charlie's facial expression and bubble (!) provide cues that the third panel is a Peak (though the event remains unknown). Nevertheless, by the Release in both Fig. 9a and b, the sequence should generate the *same* inference (Linus's retrieval of the ball).

To summarize: B&W should hypothesize that panel 3 varies from the others for Fig. 9c, while panel 4 should vary in Fig. 9d (due to *Contrast* vs. *Narration* relations). VNG would hypothesize that panel 3 in Fig. 9a would vary from the others (though all

are well-formed at this position), and would warrant a reanalysis at panel 4 because of that structural difference (a missing Peak).

The results of Cohn and Kutas (2015), a study measuring brain responses at each panel, align more with the VNG interpretation. At panel 3, the brainwaves followed their order as presented in Fig. 9 (a > b > c > d). While this brain response was interpreted as indexing a process of mental model updating (P600), it did not follow the pattern as would be assumed by B&W, where Fig. 9c would be an outlier as a *Contrast* relation compared to the other *Narration* relations (a = b = d > c). In VNG, all of these panels would be well-formed at this stage of interpretation with regard to the narrative grammar, but would differ regarding semantic changes relative to a mental model (note: 9a differed from 9b). VNG would predict that only the final panel of Fig. 9a would differ, because of the structural revision warranted by omitting a Peak panel. This was the case. An updating response (P600) occurred only for the final panel of Fig. 9a compared to all other sequence types (a > b = c = d), including Fig. 9b, where the *same inference* was generated. This again contrasts what the B&W model would expect: Fig. 9d should be the outlier as the sole *Contrast* relation compared to the other *Narration* relations. Thus, these sequences generate distinct analyses between models, and the empirical findings of panel-by-panel updating align more with VNG than a "discourse-based" interpretation.

5.3. Sequence acceptability

One of B&W's only stated "objections" with the idea of grammar is "the rigidity that such approaches appear to impose", since "naturally occurring narratives…suggest a contingent flexibility of a different order to that provided by…syntactic mechanisms" (Bateman and Wildfeuer, 2014b, p. 200). That is, all sequences are potentially construable to some form of meaning for a "defeasible interpretation," and therefore one should reject the idea of an "ungrammatical" or "ill-formed" sequence of images. However, their discussion about the acceptability of various hypothetical representations remains ungrounded in any form of empirical evidence (diagnostic or experimental), involves handwaving to justify sequences' "aesthetic" qualities, or attempts to explain away incongruities with extensive *post hoc* interpretations.

First, it is important to draw a distinction between "construability" and "well-formedness/grammaticality." Construability appears to be concerned with semantics—the ability to make meaning out of an image sequence. However, well-formedness is a judgment of the *structure* of a sequence, not necessarily its meaning. Indeed, experimental evidence has shown that participants behaviorally and neurocognitively distinguish well-formed narrative grammatical sequences, *even in the absence of semantic connections between panels* (Cohn et al., 2012).

Second, a reader can easily "construe" a sequence of images to make sense through extensive interpretation, even if it does not. B&W do this overtly when presenting the reader with the scrambled sequence in Fig. 8e, which they claim is both not "ungrammatical" yet also "certainly involves more of a challenge for finding a convincing interpretation" (Bateman and Wildfeuer, 2014a, p. 391). To explain its meaning, they offer explanations like that the scrambling of panels may "[leave] open many more issues of interpretation" and show a "flashback" or an "imagined event." B&W's interpretive offerings show that their example in fact does *not* make sense on a basic level, or else they would not need such suggestions in the first place. In contrast, assessing well-formedness is a basic, direct judgment on how an individual processes a sequence of units, and involves no such after-the-fact interpretive justification which operate after such processing is made. This difference is apparent in the timing of brain responses: recognition of semantic incongruity occurs by even 400 ms after viewing an incongruous panel (Cohn et al., 2014; Cohn et al., 2012; West and Holcomb, 2002)—far faster than any post hoc interpretive explanation can be offered.

Finally, claims of "anything goes" construability directly contrasts the experimental evidence. Several studies show that participants consciously 1) discriminate between varying degrees of comprehensible and incomprehensible manipulations to sequences and images (Cohn, 2014c; Cohn and Kutas, 2015; Cohn et al., 2012; Cohn and Wittenberg, 2015; West and Holcomb, 2002), and 2) have preferred distributions for panels in particular locations over others (Cohn, 2014c; Cohn and Wittenberg, 2015; Hagmann and Cohn, 2016). Such findings also accompany many studies showing *unconscious* responses to ungrammatical sequences (reaction times, brainwaves), beyond the awareness of participant judgments. It is worth noting that B&W's example (Fig. 8e) scrambles the panels of a sequence to appear out of order; scrambling all or some panels is one of the most basic manipulations used in experiments on sequential images, and has consistently yielded both conscious and unconscious responses of incoherence (Cohn and Wittenberg, 2015; Foulsham et al., 2016; Gernsbacher et al., 1990; Hagmann and Cohn, 2016; Inui and Miyamoto, 1981; Nagai et al., 2007; Osaka et al., 2014).

Nevertheless, VNG does not exclude the possibility that readers may construe a sequence in different ways and/or change their interpretations mid-reading. This would lead to the building of a different type of grammatical or semantic structure (as in the case of ambiguities), or would result in an updating process as a sequence is parsed in an unexpected or ill-formed way (Cohn et al., 2014; Cohn and Kutas, 2015), just as garden-path sentences or other grammatical constructions require syntactic revision (Brown and Hagoort, 2000; Osterhout and Holcomb, 1992).

5.4. Separation of grammar and meaning

The core argument made by B&W's discourse-based approach against VNG is that the structures of sequential images can be comprehended on the basis of meaningful relations between images alone, without the need for a narrative grammar. VNG agrees with the idea that meaningful relations between units must be updated as one progresses through a sequence, as

would be consistent with psychological theories of discourse (Magliano and Zacks, 2011; Zwaan and Radvansky, 1998). Empirical work has shown that discontinuity in meaning triggers a brainwave response argued to be associated with mental model updating (Donchin and Coles, 1988; Kuperberg, 2013), which in sequential images is modulated by changes in characters (Cohn and Kutas, 2017), the generation of inference (Cohn and Kutas, 2015, 2017), and incongruity to the structure of actions (Amoruso et al., 2013; Sitnikova et al., 2008), such as images with omitted or incongruous motion lines (Cohn and Maher, 2015). This updating process appears continuous and ongoing throughout a visual narrative (Cohn and Kutas, 2015; Osaka et al., 2014).

However, VNG argues that such mental model updating is not enough to account for all of sequential image comprehension. Rather, a narrative grammar is needed above and beyond the semantics of a sequence. Thus, the narrative grammar is a separate structure from meaning—operating on a different processing stream—and both contribute to comprehension (and production) of a sequence. Such a separation is supported by empirical research.

In one of the first studies of VNG, visual sequences were designed to be either coherent normal narratives or totally scrambled panels, and were contrasted with those that had a well-formed narrative structure but no meaningful relations between images (Cohn et al., 2012). These "narrative-only" sequences had a well-formed narrative arc, even though the images had no meaningful relationships to each other—analogous to Chomsky's (1965) famous sentence *Colorless green ideas sleep furiously* which has a syntactic structure, but no semantics. Participants were first shown a panel from the sequence (the target) and then pressed a button when they again saw that panel in the context of a sequence. In the first experiment, faster response times appeared to target panels within narrative-only sequences than those in scrambled sequences, but not as fast as those to panels in normal sequences. Thus, the presence of only a narrative grammar—in the absence of meaning—gave participants an advantage to their response times. A subsequent experiment then measured participants' brainwaves while viewing these same sequences. Here, a brain response typically associated with semantic processing—the "N400" (Kutas and Federmeier, 2011)—was greater to panels in scrambled and narrative-only sequences than in the normal sequences. However, the N400 did not differ between panels in scrambled and narrative-only sequences.

Thus, the brain response for semantic processing was insensitive to the presence of narrative structure, suggesting that they are different processes—even though behavioral response times benefited from the presence of this structure. If only meaningful relationships were analyzed across sequences, as argued in B&W, then there should be no behavioral advantage to the narrative-only sequences. The discourse model of B&W does not account for these experimental observations.

Additional evidence for the separation of grammar and meaning in sequential images comes from a comparison of the brainwaves observed in different studies of visual narrative processing. Over years of research, different brainwaves have been consistently observed to different neurocognitive functions. For example, the brainwave response to violations of semantics (N400) are distinctly different than those observed to violations of syntax in language (Kaan, 2007; Kuperberg, 2013). Similar brainwaves have been observed in sequential images: larger N400s appear to incongruous than congruous images within a narrative sequence (West and Holcomb, 2002) or to images within scrambled than coherent sequences (Cohn et al., 2012). In contrast, different brainwaves responses associated with the processing of syntax are evoked by violations of narrative grammar, such as the omission of a particular narrative category (Cohn and Kutas, 2015), the disruption of narrative constituents (Cohn et al., 2014), or the contrast between narrative patterns (Cohn and Kutas, 2017). Thus, different brain responses characteristic of grammar and meaning are evoked by different types of violations, as predicted by VNG.

Note also that the brainwave responses observed in these studies of sequential image processing appear to have the same characteristics as those observed in studies of language processing at the sentence level (Kaan, 2007; Kuperberg, 2013). This suggests that visual narrative sequences like those in comics are processed using similar mechanisms as in language, thereby supporting a central tenant of Visual Language Theory (Cohn, 2013b). Again, they also directly contrast B&W's claim that "very few of the properties exhibited by grammars of natural language in fact carry over to visual media such as comics" (Bateman and Wildfeuer, 2014a, p. 374). These similarities are not just at the observed structural level (via diagnostics and various experiments), but via the unconscious processing of the human brain.

5.5. Roles played by units

Because B&W's model posits a process of updating with each successive image, meaningful information is extracted from each image, and then compared with the previous one in a uniform way. However, panels themselves do not play any particular roles within a sequence. This directly contrasts with VNG's claims that panels play narrative roles relative to a broader schematic pattern. If sequences rely solely on meaningful relations, all panels within sequences should behave in the same ways. This is not the case, and panels do demonstrate unique characteristics for different roles.

In one experiment, participants were provided panels from a sequence and asked to omit one. Certain categories (Initials, Peaks) were chosen to be deleted from a sequence less often than categories that play a more "peripheral" role in the sequence (Establishers, Releases) (Cohn, 2014c). In a complementary task, participants viewed a sequence where a panel was omitted, and then identified where it had been deleted. Here, participants were more likely to recognize these same "core" categories when they are missing (Cohn, 2014c; Magliano et al., 2016). Thus, panels are not treated uniformly for their behaviors in a sequence.

In addition, some panels are more able to be reordered within in a sequence than others (Cohn, 2014c), and different brain responses appear to panels with narrative roles that are out of place in a sequence (Cohn, 2012; Cohn and Kutas, 2015). In other words, the content of some panels can adapt to various narrative roles in a sequence, while other content is more fixed.

These findings contrast the widespread assumptions upheld by B&W that *any discourse unit* can play a role in *any position* within a narrative (Sternberg, 1982). Rather, a panel can play different contextual roles depending on the particular constraints of its content, in relation to those roles imposed by a top-down structure. Again, this is similar to syntactic categories (ex. noun: *the dance* vs. verb: *they dance*). Thus, because panels of certain categories have distinguishable behaviors from others, it provides evidence that panels have varying roles from each other, and do not rely strictly on construed meaningful relations.

5.6. Structural predictions

Another aspect of B&W's model lies in its perceived directionality of comprehension. Because coherence relationships can only be derived once two images have already been read, such relations are only possible through a backward-looking process. That is, a relation can only be derived once a subsequent unit can be reached such that both units can be compared. In contrast, VNG involves both backward-looking updating *and* forward-looking predictions. In VNG, the canonical narrative schema outlines a particular expected order that types of panels will appear. Thus, when a reader is at an Initial, they have a probabilistic expectation that a Peak will follow, because Peaks follow Initials in the canonical narrative schema.

Evidence for such forward-looking predictions came from a study where blank white "disruption" panels were inserted into sequences designed to have two narrative constituents (Cohn et al., 2014). Disruptions either appeared within the constituents or at the break between the constituents with the hypothesis that violations of coherent groupings (within constituents) should be worse than those that fell between the groups. Again, a measure of brainwaves showed this outcome. At the disruption panels themselves, larger anterior negativities (a negative electrical brainwave response with a distribution across the front of the scalp) appeared to the blank panels that disrupted the constituents compared those that fell at the natural break between groupings (Cohn et al., 2014).

It is crucial to point out that these results *could not* occur if participants only used backward-looking updating of meaning. First, the observed brainwave patterns are similar to those found to violations of syntax in other domains, like language and music (Kaan, 2007; Patel, 2003). Second, a larger brainwave effect appeared to the disruptions placed *within the first constituent* relative to those between constituents. At both of these panels, participants had not yet crossed the boundary to view the panel after the constituent break. Because of this, they were unable establish a relationship between units, because they *had not yet viewed* the subsequent panel. Nevertheless, their brain responses differed between these disruptions, indicating that they made *predictions* about the narrative structure on the basis of the content, not just the relations between images. Reliance on a "panel transition" here is *impossible*. This predictive processing provides clear empirical evidence that rules out any theory of sequential images based solely on backward-looking meaningful relationships.

5.7. Stored patterns

Because B&W's discourse model stresses the "dynamic updating" of information from unit to unit, it posits no stored structures in human memory. Rather, each discourse relation is computed "dynamically" given the relational properties of compared panels. This essentially yields a uniform process of updating, characterized by different construed relations. This contrasts with VNG, which posits schematic structures stored in memory, both in the form of a canonical narrative schema, and modifiers like conjunction and Refiners.

Participants do indeed distinguish between canonical and non-canonical narrative structures (which arise in surface structures due to embedding). For example, participants are more accurate at reconstructing the order of canonical narrative sequences than those originally in non-canonical orders (Cohn, 2014c). Such stored structures are also suggested by the aforementioned studies manipulating sequences to use narrative structures but no semantic associations between panels. Participants are able to rely on stored structures of narrative without accessing a sequential meaning (Cohn et al., 2012).

Also, the lack of structures in memory does not allow for an explanation of how visual languages differ cross-culturally. VLT posits that cross-cultural variation arises because individuals within and between cultures may differ in the patterned representations they store in memory. These patterns thus manifest in distinct "visual languages" from different populations. For example, corpus analyses have suggested that the Japanese Visual Language found in manga uses E-Conjunction from the visual narrative grammar in greater proportion than American Visual Languages found in comics from the United States (Cohn, 2013b). If only dynamic construal occurred in the panel-by-panel understanding of sequential images, then readers should have no basis for such stored patterns in regularized ways across cultures. It is unclear how B&W would account for cross-cultural regularities.

Recent research has also suggested that the processing of visual narratives may vary based on the patterns found in the particular "visual language" that a person reads. In a recent study, participants' brainwaves were recorded to panels in sequences that did or did not use E-Conjunction (Cohn and Kutas, 2017). Participants evoked two neural responses. First, an anterior negativity suggested combinatorial processing to the grammatical patterning, regardless of modulation of semantic incongruity. Second, a positivity distributed in the posterior part of the scalp (a "P600") suggestive of mental model updating was sensitive to both conjunction and discontinuity in the sequence. These results were consistent with the idea that people use both a narrative grammar *and* mental model updating (for the inference of combining different characters into a common

spatial environment) in the comprehension of sequential images. They are also consistent with the waveforms observed in previous studies that violate the narrative grammar (Cohn et al., 2014).

A follow-up statistical regression analysis examined how these brain responses may have been modulated by participants' experience. Consistent with the corpus studies described above that show Japanese Visual Language uses more E-Conjunction than American comics, the only aspect of background experience that modulated the understanding of E-Conjunction was the frequency that participants read manga while growing up. When examined more closely, it was observed that frequent manga readers relied more on automatic combinatorial processing and less on updating (larger anterior negativities, smaller P600s), while infrequent manga readers did the opposite (larger P600s, smaller anterior negativities). These results were interpreted as showing that, because Japanese manga use E-Conjunction more often than American comics, readers familiar with this pattern rely more on grammatical processing, while less frequent manga readers rely more on updating a mental model. In other words, participants' brain responses differed based on the comics they read.

These results are noteworthy for two reasons. First, they show that familiarity with the patterns found in visual narratives systems modulates their understanding. This goes directly against B&W's model, which posits no structures stored in memory, and only focuses on dynamic construal of image relations. Such results are not possible with B&W's model alone. Second, though the brain response consistent with mental model updating is in line with B&W's claims of such a process, it is more pronounced in individuals who are *less familiar* with this sequencing pattern. Participants more familiar with this pattern (i.e., manga readers, as supported by corpus evidence) seem to rely more on a different neural response—one that is insensitive to modulations in the meaning of the sequence.

6. Empirical validation

Throughout B&W's papers, they emphasize the importance of the empirical verifiability of their approach, and the importance of such validation for theories.[4] They state that they "consider it essential for progress that hypotheses and analyses are pursued in a manner that is supportive of, or at least works towards, empirical evaluation" (Bateman and Wildfeuer, 2014a, p. 400). However, in the context of their papers, this claim is a promissory note, made in the absence of any actual empirical data, either provided by them or referenced from other sources. Nowhere do they provide ways for a reader to assess the validity of their own arguments. How do we know their analyses are correct? How can we test them experimentally?

By comparison, VNG not only provides the basic constructs, but also lays out diagnostic tests (Cohn, 2013c, 2014a, 2014b, 2015a), both for an analyst to probe their own interpretations, and as the basis for experimentation. In fact, analysis of sequences in VNG is not done simply by looking at a sequence and labeling things, but rather uses diagnostic tests to reveal the constructs in the first place (for a tutorial, see Cohn, 2015a). No such tests are provided by B&W, nor do they show how their own analyses fare with the diagnostics posited by VNG. B&W also do not detail how their analyses are derived—are there procedures involved, or just interpretive labeling?

In addition, despite the laudable emphasis on empirical evidence, as discussed, their model insufficiently accounts for the empirical observations available from psychological experimentation (as summarized above). Such observations go undiscussed, despite appearing in publications that B&W directly reference (Cohn, 2013b: Chapter 6). Just because their work supposedly seeks to go beyond the psycholinguistic approach (Bateman and Wildfeuer, 2014a, p. 400), their own model is not exempt from accounting for the existing empirical evidence provided by actual experiments on cognition.

To take a strong position, evidence from empirical experimentation is the *only* way to validate a theory, because all models are supposedly characterizing the way that the brain processes this information. If a model attempts to characterize how people actually comprehend a sequence of images (or any other facet of the visual language in comics), it must be testable and validated by psychological experimentation. If a model is not meant to characterize cognition (e.g., a literary theory), then it need not live up to this standard, but it should then not be attempted to be compared to theories that actually do.

Thus, while this review has targeted the approach by B&W specifically, it serves as a broader caveat to the growing endeavors investigating comics, in line with previous advocacy (Cohn, 2014b). Theoretical claims should be backed by data, or at least should articulate empirically testable hypotheses and formulations. Such theories and/or empirical research should also reference and be supported by the existing literature. These standards of research should be upheld by researchers wishing to make claims about the constructs underlying comics, whether they support or refute the language-like qualities of such structures.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.pragma.2018.01.002.

[4] It should be noted that this emphasis has laudably been followed up by efforts to bring together scholars interested in empirically driven comics research in conferences and book collections. Thus, B&W legitimately are committed to empirical validity, and their statements should not be taken as lip service.

References

Amoruso, L., Gelormini, C., Aboitiz, F., Alvarez González, M., Manes, F., Cardona, J., Ibanez, A., 2013. N400 ERPs for actions: building meaning in context. Front. Hum. Neurosci. 7 https://doi.org/10.3389/fnhum.2013.00057.

Asher, N., Lascarides, A., 2003. Logics of Conversation. Cambridge University Press, Cambridge.

Bateman, J.A., Wildfeuer, J., 2014a. Defining units of analysis for the systematic analysis of comics: a discourse-based approach. Stud. Comics 5 (2), 373–403. https://doi.org/10.1386/stic.5.2.373_1.

Bateman, J.A., Wildfeuer, J., 2014b. A multimodal discourse theory of visual narrative. J. Pragmat. 74, 180–208. https://doi.org/10.1016/j.pragma.2014.10.001.

Brown, C.M., Hagoort, P., 2000. On the electrophysiology of language comprehension: implications for the human language system. In: Crocker, M.W., Pickering, M.J., Clifton Jr., C. (Eds.), Architectures and Mechanisms for Language Processing. Cambridge University Press, Cambridge, UK, pp. 213–237.

Carroll, J.M., 1980. Toward a Structural Psychology of Cinema. Mouton, The Hague.

Cheng, L.L.-S., Corver, N., 2013. Diagnosing Syntax. Oxford University Press.

Chomsky, N., 1965. Aspects of the Theory of Syntax. MIT Press, Cambridge, MA.

Chomsky, N., 1972. Studies on Semantics in Generative Grammar. Mouton, The Hague.

Chomsky, N., 1981. Lectures on Government and Binding. Foris, Dordrecht.

Cohn, N., 2003. Early Writings on Visual Language. Emaki Productions, Carlsbad, CA.

Cohn, N., 2010. The limits of time and transitions: challenges to theories of sequential image comprehension. Stud. Comics 1 (1), 127–147.

Cohn, N., 2012. Structure, Meaning, and Constituency in Visual Narrative Comprehension (Doctoral Dissertation). Tufts University, Medford, MA.

Cohn, N., 2013a. Navigating comics: an empirical and theoretical approach to strategies of reading comic page layouts. Front. Psychol. Cogn. Sci. 4, 1–15. https://doi.org/10.3389/fpsyg.2013.00186.

Cohn, N., 2013b. The Visual Language of Comics: Introduction to the Structure and Cognition of Sequential Images. Bloomsbury, London, UK.

Cohn, N., 2013c. Visual narrative structure. Cogn. Sci. 37 (3), 413–452. https://doi.org/10.1111/cogs.12016.

Cohn, N., 2014a. The architecture of visual narrative comprehension: the interaction of narrative structure and page layout in understanding comics. Front. Psychol. 5, 1–9. https://doi.org/10.3389/fpsyg.2014.00680.

Cohn, N., 2014b. Building a better "comic theory": shortcomings of theoretical research on comics and how to overcome them. Stud. Comics 5 (1), 57–75. https://doi.org/10.1386/stic.5.1.57_1.

Cohn, N., 2014c. You're a good structure, Charlie Brown: the distribution of narrative categories in comic strips. Cogn. Sci. 38 (7), 1317–1359. https://doi.org/10.1111/cogs.12116.

Cohn, N., 2015a. How to Analyze Visual Narratives: a Tutorial in Visual Narrative Grammar. Retrieved from. http://www.visuallanguagelab.com/P/VNG_Tutorial.pdf.

Cohn, N., 2015b. Narrative conjunction's junction function: the interface of narrative grammar and semantics in sequential images. J. Pragmat. 88, 105–132. https://doi.org/10.1016/j.pragma.2015.09.001.

Cohn, N., 2016. A multimodal parallel architecture: a cognitive framework for multimodal interactions. Cognition 146, 304–323. https://doi.org/10.1016/j.cognition.2015.10.007.

Cohn, N., Campbell, H., 2015. Navigating comics II: constraints on the reading order of page layouts. Appl. Cogn. Psychol. 29 (2), 193–199. https://doi.org/10.1002/acp.3086.

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. https://doi.org/10.1016/j.neuropsychologia.2014.09.018.

Cohn, N., Kutas, M., 2015. Getting a cue before getting a clue: event-related potentials to inference in visual narrative comprehension. Neuropsychologia 77, 267–278. https://doi.org/10.1016/j.neuropsychologia.2015.08.026.

Cohn, N., Kutas, M., 2017. What's your neural function, visual narrative conjunction? Grammar, meaning, and fluency in sequential image processing. Cogn. Res. Princ. .Implic. 2 (27), 1–13. https://doi.org/10.1186/s41235-017-0064-5.

Cohn, N., Maher, S., 2015. The notion of the motion: the neurocognition of motion lines in visual narratives. Brain Res. 1601, 73–84. https://doi.org/10.1016/j.brainres.2015.01.018.

Cohn, N., Murthy, B., Foulsham, T., 2016. Meaning above the head: combinatorial constraints on the visual vocabulary of comics. J. Cogn. Psychol. 28 (5), 559–574. https://doi.org/10.1080/20445911.2016.1179314.

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. Cogn. Psychol. 65 (1), 1–38. https://doi.org/10.1016/j.cogpsych.2012.01.003.

Cohn, N., Wittenberg, E., 2015. Action starring narratives and events: structure and inference in visual narrative comprehension. J. Cogn. Psychol. 27 (7), 812–828. https://doi.org/10.1080/20445911.2015.1051535.

Culicover, P.W., Jackendoff, R., 2005. Simpler Syntax. Oxford University Press, Oxford.

Donchin, E., Coles, M.G.H., 1988. Is the P300 component a manifestation of context updating? Behav. Brain Sci. 11 (03), 357–374. https://doi.org/10.1017/S0140525X00058027.

Foulsham, T., Wybrow, D., Cohn, N., 2016. Reading without words: eye movements in the comprehension of comic strips. Appl. Cogn. Psychol. 30, 566–579. https://doi.org/10.1002/acp.3229.

Gernsbacher, M.A., 1985. Surface information loss in comprehension. Cogn. Psychol. 17, 324–363.

Gernsbacher, M.A., Varner, K.R., Faust, M., 1990. Investigating differences in general comprehension skill. J. Exp. Psychol. Learn. Mem. Cogn. 16, 430–445.

Goldberg, A., 1995. Constructions: a Construction Grammar Approach to Argument Structure. University of Chicago Press, Chicago, IL.

Hagmann, C.E., Cohn, N., 2016. The pieces fit: constituent structure and global coherence of visual narrative in RSVP. Acta Psychol. 164, 157–164. https://doi.org/10.1016/j.actapsy.2016.01.011.

Inui, T., Miyamoto, K., 1981. The time needed to judge the order of a meaningful string of pictures. J. Exp. Psychol. Hum. Learn. Mem. 7 (5), 393–396.

Jackendoff, R., 1990. Semantic Structures. MIT Press, Cambridge, MA.

Jackendoff, R., 2002. Foundations of Language: Brain, Meaning, Grammar, Evolution. Oxford University Press, Oxford.

Kaan, E., 2007. Event-related potentials and language processing: a brief overview. Lang. Ling. Compass 1 (6), 571–591. https://doi.org/10.1111/j.1749-818X.2007.00037.x.

Kuperberg, G.R., 2013. The pro-active comprehender: what event-related potentials tell us about the dynamics of reading comprehension. In: Miller, B., Cutting, L., McCardle, P. (Eds.), Unraveling the Behavioral, Neurobiological, and Genetic Components of Reading Comprehension. Paul Brookes Publishing, Baltimore, pp. 176–192.

Kutas, M., Federmeier, K.D., 2011. Thirty years and counting: finding meaning in the N400 component of the Event-Related Brain Potential (ERP). Annu. Rev. Psychol. 62 (1), 621–647.

Ludlow, P., 2011. The Philosophy of Generative Linguistics. Oxford University Press, Oxford, UK.

Magliano, J.P., Kopp, K., Higgs, K., Rapp, D.N., 2016. Filling in the gaps: memory implications for inferring missing content in graphic narratives. Discourse Process. https://doi.org/10.1080/0163853X.2015.1136870, 0–0.

Magliano, J.P., Larson, A.M., Higgs, K., Loschky, L.C., 2015. The relative roles of visuospatial and linguistic working memory systems in generating inferences during visual narrative comprehension. Mem. Cogn. 44 (2), 207–219. https://doi.org/10.3758/s13421-015-0558-7.

Magliano, J.P., Miller, J., Zwaan, R.A., 2001. Indexing space and time in film understanding. Appl. Cogn. Psychol. 15, 533–545.

Magliano, J.P., Zacks, J.M., 2011. The impact of continuity editing in narrative film on event segmentation. Cogn. Sci. 35 (8), 1489–1517. https://doi.org/10.1111/j.1551-6709.2011.01202.x.

Mandler, J.M., Johnson, N.S., 1977. Remembrance of things parsed: story structure and recall. Cogn. Psychol. 9, 111–151.

McCloud, S., 1993. Understanding Comics: the Invisible Art. Harper Collins, New York, NY.
Nagai, M., Endo, N., Takatsune, K., 2007. Measuring brain activities related to understanding using near-infrared spectroscopy (NIRS). In: Smith, M.J., S. G (Eds.), Human Interface and the Management of Information: Methods, Techniques and Tools in Information Design, vol. 4557. Springer Berlin, Heidelberg, pp. 884–893.
Omori, T., Ishii, T., Kurata, K., 2004. Eye Catchers in Comics: Controlling Eye Movements in Reading Pictorial and Textual Media. Paper presented at the 28th International Congress of Psychology, Beijing, China. http://www.cirm.keio.ac.jp/media/contents/2004ohmori.pdf.
Osaka, M., Yaoi, K., Minamoto, T., Osaka, N., 2014. Serial changes of humor comprehension for four-frame comic Manga: an fMRI study. Sci. Rep. 4 https://doi.org/10.1038/srep05828.
Osterhout, L., Holcomb, P., 1992. Event-related potentials elicited by syntactic anomaly. J. Mem. Lang. 31, 758–806.
Patel, A.D., 2003. Language, music, syntax and the brain. Nat. Neurosci. 6 (7), 674–681. https://doi.org/10.1038/nn1082.
Postal, P.M., 1972. The best theory. In: Peters, R.S. (Ed.), Goals of Linguistic Theory. Prentice-Hall, Englewood Cliffs, NJ, pp. 131–170.
Saraceni, M., 2000. Language beyond Language: Comics as Verbo-visual Texts (Doctoral Dissertation). University of Nottingham, Nottingham.
Saraceni, M., 2016. Relatedness: aspects of textual connectivity in comics. In: Cohn, N. (Ed.), The Visual Narrative Reader. Bloomsbury, London, pp. 115–129.
Sitnikova, T., Holcomb, P.J., Kuperberg, G.R., 2008. Two neurocognitive mechanisms of semantic integration during the comprehension of visual real-world events. J. Cogn. Neurosci. 20 (11), 1–21.
Stainbrook, E.J., 2003. Reading Comics: a Theoretical Analysis of Textuality and Discourse in the Comics Medium (Doctoral Dissertation). Indiana University of Pennsylvania, Indiana, PA.
Stainbrook, E.J., 2016. A little cohesion between friends; or, We're just exploring our textuality: reconciling cohesion in written language and visual language. In: Cohn, N. (Ed.), The Visual Narrative Reader. Bloomsbury, London, pp. 129–154.
Sternberg, M., 1982. Proteus in quotation-land: Mimesis and the forms of reported discourse. Poetics Today 3 (2), 107–156.
West, W.C., Holcomb, P., 2002. Event-related potentials during discourse-level semantic integration of complex pictures. Cogn. Brain Res. 13, 363–375.
Zwaan, R.A., Radvansky, G.A., 1998. Situation models in language comprehension and memory. Psychol. Bull. 123 (2), 162–185.

Neil Cohn is an assistant professor at the Tilburg center for Cognition and Communication at Tilburg University. He is internationally recognized for his research on the overlap of the structure and cognition of sequential images and language. His books *The Visual Language of Comics* (Bloomsbury, 2013) and *The Visual Narrative Reader* (Bloomsbury, 2016) integrate interdisciplinary research on visual narratives into a unified field within the linguistic and cognitive sciences. His work is online at www.visuallanguagelab.com.