Neuropsychologia 64 (2014) 63–70

Contents lists available at ScienceDirect

Neuropsychologia

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

The grammar of visual narrative: Neural evidence for constituent structure in sequential image comprehension

Neil Cohn [a,c,*], Ray Jackendoff [b,a], Phillip J. Holcomb [a], Gina R. Kuperberg [a,d]

[a] Department of Psychology, Tufts University, 490 Boston Ave, Medford, MA 02155, USA
[b] Department of Philosophy and Center for Cognitive Studies, Tufts University, 115 Miner Hall, Medford, MA 02155, USA
[c] Department of Cognitive Science, University of California, San Diego, 9500 Gilman Dr. Dept 0526, La Jolla, CA 92093-0526, USA
[d] Department of Psychiatry and Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Bldg 149, 13th Street, Charlestown, MA 02129, USA

ARTICLE INFO

Article history:
Received 8 June 2014
Received in revised form
26 August 2014
Accepted 10 September 2014
Available online 18 September 2014

Keywords:
Constituent structure
Grammar
Narrative
Visual language
Comics
ERPs

ABSTRACT

Constituent structure has long been established as a central feature of human language. Analogous to how syntax organizes words in sentences, a narrative grammar organizes sequential images into hierarchic constituents. Here we show that the brain draws upon this constituent structure to comprehend wordless visual narratives. We recorded neural responses as participants viewed sequences of visual images (comics strips) in which blank images either disrupted individual narrative constituents or fell at natural constituent boundaries. A disruption of either the first or the second narrative constituent produced a left-lateralized anterior negativity effect between 500 and 700 ms. Disruption of the second constituent also elicited a posteriorly-distributed positivity (P600) effect. These neural responses are similar to those associated with structural violations in language and music. These findings provide evidence that comprehenders use a narrative structure to comprehend visual sequences and that the brain engages similar neurocognitive mechanisms to build structure across multiple domains.

1. Introduction

Constituent structure is a hallmark of human language. Discrete units (words) group into larger constituents (phrases), which can recursively combine in indefinitely many ways (Chomsky, 1965; Culicover & Jackendoff, 2005). Language, however, is not our only means of communication. For millennia, humans have told stories using sequential images, whether on cave walls or paintings, or in contemporary society, in comics or films (Kunzle, 1973; McCloud, 1993). Analogous to the way words combine in language, individual images can combine to form larger constituents that enable the production and comprehension of complex coherent visual narratives (Carroll & Bever, 1976; Cohn, 2013b; Cohn, Paczynski, Jackendoff, Holcomb, & Kuperberg, 2012; Gernsbacher, 1985; Zacks, Speer, & Reynolds, 2009).

It has long been recognized that narratives follow a particular structure (Freytag, 1894; Mandler & Johnson, 1977). This dates all the way back to Aristotle's observations about plot structure in theater (Butcher, 1902). We have recently formalized a narrative grammar of sequential images, in which each image plays a *categorical role* based on its narrative function within the overall visual sequence (Cohn, 2013b, 2014). These image units can subsequently group together to form *narrative constituents*, which themselves fulfill narrative roles in the overall structure. While this general approach is similar to previous grammars of discourse and stories (e.g., Clark, 1996; Hinds, 1976; Labov & Waletzky, 1967; Mandler & Johnson, 1977; Rumelhart, 1975), it differs from these precedents in the simplicity of its recursive structures (Cohn, 2013b), the incorporation of modifiers beyond a canonical narrative arc (Cohn, 2013a, 2013b), and the explicit separation of structure and meaning (Cohn et al., 2012), see Cohn (2013b) for more details.

To better understand this narrative grammar, consider the sequence shown in Fig. 1. This has two narrative constituents. The first constituent contains two images: the first image plays the narrative role of an "Initial," functioning to set up the central event ("hitting the ball"), while the second image plays the narrative role of a "Peak" as it depicts the hitting action itself. The second constituent consists of four images: the first, an "Establisher," functions to introduce the characters involved in the main event; the second, an "Initial," sets up the event; the third functions as a "Peak" depicting the climactic crashing event itself, and the fourth image acts as a "Release," resolving this central action. Importantly, these two narrative constituents are related on a higher level of narrative structure, such that the first larger constituent

* Corresponding author at: Department of Cognitive Science, University of California, San Diego, 9500 Gilman Dr. Dept. 0526, La Jolla, CA 92093-0526, USA. Tel.: +1 858 822 0736; fax: +1 858 822 5097.
E-mail address: neilcohn@visuallanguagelab.com (N. Cohn).

http://dx.doi.org/10.1016/j.neuropsychologia.2014.09.018
0028-3932/

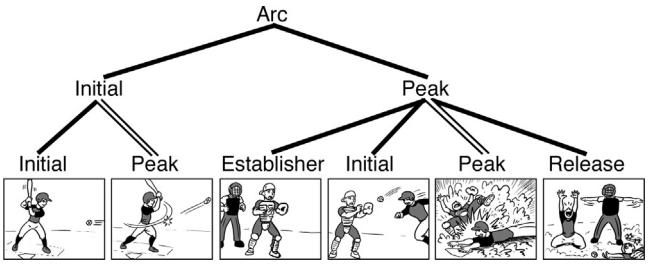

Fig. 1. Narrative structure of a visual sequence. This sequence contains two narrative constituents. The first two panels together depict an event; the first panel is an "Initial," which sets up the climatic event in the second panel, a "Peak." In turn, these two panels together serve as an Initial for the sequence as a whole, and the event depicted by the final four panels serves as Peak for the entire sequence.

functions as an "Initial" to set up the content of the second larger constituent, which itself acts as the climactic "Peak" of the whole sequence. In more complex narratives, embedding along similar lines can be deeper or altered through modifiers.

In previous experimental research, we have shown that these narrative categories follow distributional trends in sequences, relying on cues from both image content as well as their context within a sequence (Cohn, 2014). In addition, our previous work suggests that, during the comprehension of visual narrative sequences, the brain uses this narrative structure in combination with more general semantic schemas (Schank & Abelson, 1977) to build up global narrative coherence, which, in turn, facilitates semantic processing of incoming panels (Cohn et al., 2012). So far, however, it remains unclear how the brain responds to input that actually violates expectations that are based on our representation of this narrative structure. Addressing this question was the aim of the present study. We show that the constituent structure in our proposed narrative grammar is not just an interesting theoretical construct: it can be detected experimentally.

The paradigm we developed is modeled on classic psycholinguistic experiments that demonstrated that word-by-word comprehension engages grammatical constituent structure. In an important series of behavioral studies, participants listened to simple sentences such as *My roommate watched the television*, during which there was a burst of white noise (a "click": depicted here as **). Initial research using this paradigm showed that clicks appearing *within* a syntactic constituent (e.g., disrupting the noun-phrase: *My ** roommate watched…*) were recalled less accurately than clicks appearing *between* syntactic constituents (e.g., between the noun-phrase and the verb-phrase: *My roommate ** watched…*), and that false recollection of clicks remembered them as occurring between constituents (Fodor & Bever, 1965; Garrett & Bever, 1974). Later studies using online monitoring tasks found that reaction times were faster to clicks placed between constituents than those within syntactic constituents, and faster to those within first constituents than second constituents (Abrams & Bever, 1969; Bond, 1972; Ford & Holmes, 1978). The success of this "structural disruption" technique as a method of examining grammatical structure in language has led to its use beyond the study of structure in language, to study structure in music (Berent & Perfetti, 1993; Kung, Tzeng, Hung, & Wu, 2011) and visual events (Baird & Baldwin, 2001).

At a neural level, studies using event-related potentials (ERPs) have reported two effects in association with structural (syntactic) aspects of language processing: (1) a left-lateralized anterior negativity (Friederici, 2002; Friederici, Pfeifer, & Hahne, 1993; Hagoort, 2003; Neville, Nicol, Barss, Forster, & Garrett, 1991), starting at or before 350 ms, and (2) a posteriorly-distributed positivity (P600), starting at around 500 ms, although sometimes earlier (Hagoort, Brown, & Groothusen, 1993; Osterhout & Holcomb, 1992). The left anterior negativity effect tends to be evoked by words that are consistent (versus inconsistent) with one of just two or three possible upcoming syntactic structures predicted by the context (Lau, Stroud, Plesch, & Phillips, 2006), and it is seen even when this context is semantically non-constraining (Gunter, Friederici, & Schriefers, 2000) or semantically incoherent (Münte, Matzke, & Johannes, 1997). The P600 is most likely to be triggered when an input violates a strong, high certainty single structural expectation established by a context, particularly when this context is also semantically constraining (Kuperberg, 2007, 2013). It is believed to reflect prolonged attempts to make sense of the input (Kuperberg, 2007, 2013; Sitnikova, Holcomb, & Kuperberg, 2008a). Notably, both the left anterior negativity and P600 effects are distinct from the well-known N400 effect––a widespread negativity between 300 and 500 ms that is modulated by both words (Kutas & Hillyard, 1980) and images (Barrett & Rugg, 1990; Barrett, Rugg, & Perrett, 1988) that match versus mismatch contextual expectations about the semantic features of upcoming input, rather than expectations about its grammatical structure (Kutas & Federmeier, 2011).

Here, we ask whether violations of narrative constituent structure in sequential images produce neural effects analogous to those seen in response to structural violations in language. We developed a structural disruption paradigm, analogous to the classic "click" paradigm that, as discussed, provided early evidence that comprehenders use a syntactic constituent structure to comprehend language (Fodor & Bever, 1965; Garrett & Bever, 1974). Our paradigm also shares similarities with so-called ERP "omission" paradigms in which, rather than examining the neural response to a stimulus that is incongruous (versus congruous) with a context, ERPs are time-locked to the *omission* of the expected stimulus. We have known since the late 1960s that the omission of expected stimuli can evoke a large brain response (Klinke, Fruhstorfer, & Finkenzeller, 1968; Simson, Vaughan, & Walter, 1976), and more recently, this phenomenon has been interpreted within a generative Bayesian predictive coding framework (Clark, 2013; Friston, 2005; Rao & Ballard, 1999). According to this framework, the brain constructs an internal model of the environment by constantly assessing incoming stimuli in relation to their preceding context and stored representations. Top-down predictions are compared, at multiple levels of representation, with incoming stimuli, and the difference in the neural response between the top-down prediction and the bottom-up input—the "prediction error"—is passed up to a higher level of representation, where it is used to adjust the internal model or, when the input violates a very high certainty expectation, switch to an alternative model that can better explain the combination of the context and the incoming stimulus. Neural responses to omissions are taken to reflect pure neural prediction error, produced by the mismatch between top-down predictions and (absent) bottom-up input (Bendixen, Schröger, & Winkler, 2009; Friston, 2005; Todorovic, van Ede, Maris, & de Lange, 2011; Wacongne et al., 2011).

In our paradigm, participants viewed six-panel-long wordless visual sequences, presented image-by-image. These panels were constructed to have two narrative constituents in various different structural patterns (see Section 2). In some of the visual sequences, we inserted "blank" white panels devoid of content ("omission" stimuli). The blank panels fell either *within* a narrative constituent (either the first or the second constituent) or *in between* the two narrative constituents (see Fig. 2 for an example). Importantly, because we used several patterns of constituent structure, with narrative boundaries located after panel 2, 3, or 4, blank panels could appear anywhere from the second to fifth panel position in the sequence. This meant that comprehenders could not use ordinal position as a direct cue to predict when a blank panel would occur. We measured ERPs to these blank panels, and compared the ERP response produced by those that fell *within*