Cognition 146 (2016) 304–323

Contents lists available at ScienceDirect

Cognition

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

A multimodal parallel architecture: A cognitive framework for multimodal interactions

Neil Cohn *

Center for Research in Language, University of California, San Diego, United States

ARTICLE INFO

Article history:
Received 20 August 2014
Revised 3 September 2015
Accepted 11 October 2015
Available online 9 November 2015

Keywords:
Multimodality
Visual language
Gesture
Comics
Narrative
Parallel architecture
Linguistic models

ABSTRACT

Human communication is naturally multimodal, and substantial focus has examined the semantic correspondences in speech–gesture and text–image relationships. However, visual narratives, like those in comics, provide an interesting challenge to multimodal communication because the words and/or images can guide the overall meaning, and both modalities can appear in complicated "grammatical" sequences: sentences use a syntactic structure and sequential images use a narrative structure. These dual structures create complexity beyond those typically addressed by theories of multimodality where only a single form uses combinatorial structure, and also poses challenges for models of the linguistic system that focus on single modalities. This paper outlines a broad theoretical framework for multimodal interactions by expanding on Jackendoff's (2002) parallel architecture for language. Multimodal interactions are characterized in terms of their component cognitive structures: whether a particular modality (verbal, bodily, visual) is present, whether it uses a grammatical structure (syntax, narrative), and whether it "dominates" the semantics of the overall expression. Altogether, this approach integrates multimodal interactions into an existing framework of language and cognition, and characterizes interactions between varying complexity in the verbal, bodily, and graphic domains. The resulting theoretical model presents an expanded consideration of the boundaries of the "linguistic" system and its involvement in multimodal interactions, with a framework that can benefit research on corpus analyses, experimentation, and the educational benefits of multimodality.

1. Introduction

Humans communicate through different modalities—whether through speech, bodily movements, or drawings—and can combine these expressive capacities together in rich and complex ways. Researchers have long shown that co-speech gesture enriches communication beyond speech alone (Clark, 1996; Goldin-Meadow, 1999, 2003a; McNeill, 1992, 2000b), and growing research has investigated the various interactions between text and images (for review, see Bateman, 2014; e.g., Kress, 2009; Kress & van Leeuwen, 2001; Mayer, 2009; Mitchell, 1986). These works often examine multimodal interactions where only a single modality uses combinatorial structure across a sequence, such as using sentences (with a syntactic structure) in combination with gestures or single images (without a grammar). Yet, visual narratives in works such as comics often combine written language with a "visual language" of images (Cohn, 2013b) to create complex interactions involving both the grammar of sequential words (syntax) and the grammar of sequential images (narrative structure) as the dual packagers of meaning. Such structure yields complexity beyond that typically shown in co-speech gestures or the binding of text with individual images (Cohn, 2013a).

This work seeks to characterize such complex multimodal interactions by expanding on Jackendoff's (2002) *parallel architecture* for language. Here, focus will be placed on how grammar and meaning coalesce in multimodal interactions, extending beyond the semantic taxonomies typically discussed about text–image relations (e.g., Kress, 2009; Martinec & Salway, 2005; McCloud, 1993; Royce, 2007). While work on co-speech gesture has begun to incorporate grammar into multimodal models (Fricke, 2013), the presence of "grammar" concurrently in multiple modalities poses new challenges. Moreover, most approaches to text–image relations make little attempt to integrate their observations with models of language or cognition (e.g., Kress, 2009; Martinec & Salway, 2005; McCloud, 1993; Painter, Martin, & Unsworth, 2012; Royce, 2007), or do so in ways that are insensitive to the internal structures of each modality's expressions (e.g., Mayer, 2009). Though the primary focus will remain on drawn visual

* Address: Center for Research in Language, University of California, San Diego, 9500 Gilman Dr. Dept. 0526, La Jolla, CA 92093-0526, United States.
E-mail address: neilcohn@visuallanguagelab.com

http://dx.doi.org/10.1016/j.cognition.2015.10.007
0010-0277/

narratives, by examining these complex structures, this approach can subsume aspects of co-gesture and other text–image interactions. The model arising from this approach can frame an expanded consideration of the boundaries of the "linguistic" system and its involvement in multimodal interactions, while also providing a framework that can benefit corpus analyses, experimentation, and research on the educational benefits of multimodality (Goldin-Meadow, 2003a; Mayer, 2005, 2009).

The multimodal interactions described in this work will be supported by manipulating multimodal "utterances" through diagnostic tests of deletion (omission of elements) and substitution (replacement of elements), and readers will be asked to rely on their intuitions to assess their felicity. This methodology has been common in theoretical linguistic research for decades, though criticized by some (e.g., Gibson & Fedorenko, 2010) while defended by others (e.g., Culicover & Jackendoff, 2010). Ultimately, this overall research program extends beyond intuitive judgments, and these theoretical constructs can frame empirical experimentation and corpus analyses that can validate, clarify, and/or alter the theory, much as observations from linguistics have framed psycholinguistics research. Such a research program has already been successful in studying visual narratives, where theoretical diagnostics (Cohn, 2013c, 2014a) provide the basis for experimental designs (Cohn, 2014b; Cohn, Jackendoff, Holcomb, & Kuperberg, 2014; Cohn, Paczynski, Jackendoff, Holcomb, & Kuperberg, 2012; Cohn & Wittenberg, 2015) which in turn inform the theory.

The investigation of multimodal interactions is complex. All non-attributed images have thus been created as exemplars for demonstrating the dimensions of this model as clearly as possible. However, it is fully acknowledged that "attested"[1] instances of visual narratives from comics and other domains are more complicated, and the final section provides tools for analyzing such examples using this model.

1.1. Multimodal semantic interactions

Many theoretical approaches have characterized the multimodal interactions between written and visual information (Bateman, 2014). Most of these approaches focus on the physical or semantic relationships between modalities (Forceville & Urios-Aparisi, 2009; Hagan, 2007; Horn, 1998; Kress, 2009; Martinec & Salway, 2005; McCloud, 1993; Painter et al., 2012; Royce, 2007), the socio-semiotic interpretations resulting from such interactions (Kress, 2009; Kress & van Leeuwen, 2001; Royce, 1998, 2007), and/or the benefits of multimodal relations for learning (Ayres & Sweller, 2005; Mayer, 2005, 2009). For example, Martinec and Salway (2005) describe how text or images may elaborate, extend, or enhance the meaning across modalities, while Royce (2007) characterizes traditional linguistic relations like modalities conveying the same (synonymy) or different (antonymy) meanings, crossing taxonomic levels (hyponymy), and part-whole relations (meronymy), among others. By focusing on the *semantic* aspects of text–image relationships, such approaches are commensurate with research detailing the ways that gestures match or mismatch the content of speech (e.g., Goldin-Meadow, 2003a).

Similar semantic analyses appear for multimodality in drawn visual narratives specifically. For example, Painter et al. (2012) outlined several socio-semiotic functions of interpreting text and image in children's picture books, while Bateman and Wildfeuer (2014) incorporate multimodal relations into a general framework for uniformly describing discourse relations of all sequential images. Stainbrook (2003, 2015) meanwhile has argued that consistent surface coherence relations maintain between images, text, and their relations in visual narratives. Finally, the most popularly-known approach to visual narrative multimodality comes in McCloud's (1993) broad characterization for the semantic contributions of text and image in comics. Let's examine his seven categories of "text–image" relationships more closely:

1. *Word-Specific* – Pictures illustrate but do not significantly add to the meaning given by the text.
2. *Picture-Specific* – Words only provide a "soundtrack" to a visually told sequence.
3. *Duo-Specific* – Both words and pictures send the same message.
4. *Additive* – One form amplifies or elaborates on the other.
5. *Parallel* – Words and images follow non-intersecting semantic discourses.
6. *Interdependent* – Both modalities combine to create an idea beyond the scope of either on their own.
7. *Montage* – Words are treated as part of the image itself.

This approach does not detail specific semantic relations between modalities, as found in other approaches. Rather, this taxonomy outlines a graded exchange of meaning between modalities (*Picture-Specific* to *Word-Specific*), along with several interactions where each modality has equal weight. McCloud's proposal also fits his approach to sequential image comprehension, which posits that readers generate inferences between all panel juxtapositions. This theory resembles work in discourse that details the semantic relations between sentences (e.g., Halliday & Hasan, 1976; Hobbs, 1985; Kehler, 2002; Zwaan & Radvansky, 1998). While not stated explicitly, McCloud's overall approach implies that panels create a "text–image unit," which then engages in a semantic relationship with each subsequent text–image unit.

Though this model provides a foundation for varying text–image relationships, McCloud's approach (and others) cannot account for certain contrasts between multimodal interactions. Consider Fig. 1a and b, which both might be characterized as Word-Specific in McCloud's taxonomy, since the text carries more weight of the meaning. We can test this "semantic dominance" by deleting the text from each sequence (Fig. 1c and d). In both, the overall multimodal meaning is lost: the sequences no longer convey their original meanings. While omitting the text makes both harder to understand (since the dominant carrier of meaning is gone), the isolated visual sequence in Fig. 1a makes no sense (Fig. 1c), but omitting the text in Fig. 1b retains some coherence between panels (Fig. 1d). Thus, these sequences vary in ways that McCloud's approach cannot characterize, namely multimodal interactions where the properties of the visual narrative *sequence* differ.

1.2. Structure and meaning in visual narratives

This limitation of McCloud's multimodal approach aligns with deficiencies in his model of sequential image comprehension, which focuses on changes in linear semantic coherence relationships (Cohn, 2010b, 2013c). Fig. 2 depicts a narrative sequence from Stan Sakai's *Usagi Yojimbo* that illustrates several problems with a strictly semantic approach to sequential images. Here, a ninja (in black, panel 1) uses a ball and chain to hold the sword of a samurai (the rabbit, panel 2), until the ninja jumps (panel 3) and the rabbit draws his sword (panel 4), culminating in the samurai cutting down the ninja (panel 5).

First, connections between panels extend beyond linear relationships, and could possibly span distances in a sequence (i.e., distance dependencies). In Fig. 2, panel 1 logically should connect with 3 and 5, while panel 2 must connect with 4 and 5, because the same characters repeat in those panels. Second, despite these distant relationships, we can recognize that pairs of

[1] It should be noted that, even though these examples are created for this particular context, as I am a "fluent speaker" of this visual language, these constructed examples are still "naturalistic" instances of multimodal interactions.